Estimation and Detection with Chaotic Systems

Michael D. Richard

RLE Technical Report No. 581

February 1994

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307

This work was supported in part by the U.S. Air Force Office of Scientific Research under the Augmentation Awards for Science and Engineering Research Training Program Grant F49620-92-J-0255, in part by the U.S. Air Force Office of Scientific Research Grant AFOSR-91-0034-C, in part by the U.S. Navy Office of Naval Research under Grant N00014-93-1-0686, and in part by a subcontract from Lockheed Sanders, Inc. under the U.S. Navy Office of Naval Research Contract N00014-91-C-0125.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
FEB 1994		**00-02-1994 to 00-02-1994**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Estimation and Detection with Chaotic Systems	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Laboratory for Information and Decision Systems,77 Massachusetts Avenue,Cambridge,MA,02139-4307**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**214**	

Estimation and Detection with Chaotic Systems

by

Michael D. Richard

Submitted to the Department of Electrical Engineering and Computer Science
on 1 November 1993, in partial fulfillment of the
requirements for the degree of
Doctor of Science

Abstract

Chaotic systems have received much attention in the mathematics and physics communities in the last two decades; and they are receiving increasing attention in various engineering disciplines as well. Experimental evidence suggests that these systems may be useful models for a wide variety of physical phenomena, including turbulence, vibrations of buckled elastic structures, and behavior of certain feedback control devices.

This thesis deals with both the analysis and synthesis of chaotic maps and time-sampled chaotic flows, with a focus on the problems and issues that arise with noise-corrupted orbit segments generated by these maps and flows. Both dissipative and nondissipative systems are considered, with both types of systems considered in the context of analysis and the latter type also considered in the context of synthesis. With respect to dissipative systems, three probabilistic state estimation algorithms are introduced and applied to three problem scenarios, with the scenarios distinguished by the amount of *a priori* knowledge of the dynamics of the underlying chaotic system.

Cramer-Rao, Barankin, and Weiss-Weinstein upper bounds on state estimator performance are derived and both experimentally and qualitatively analyzed. The analysis reveals that intrinsic properties of chaotic systems—positive Lyapunov exponents and boundedness of attractors—have a fundamental influence on achievable state estimator performance with these systems.

With respect to nondissipative systems, the thesis considers a class of piecewise linear maps of the unit interval, members of which give rise to finite-state, homogeneous Markov chains. The thesis establishes ergodic and other properties of these maps and explores the use of these maps as generators of signals for practical applications. A close relation is established between noise-corrupted orbit segments generated by the maps and outputs of hidden Markov models, and this relation is exploited in practical, optimal and suboptimal algorithms for detection, parameter estimation, and state estimation with the maps.

Thesis Supervisor: Alan V. Oppenheim
Title: Distinguished Professor of Electrical Engineering

Acknowledgments

As I look back over the last five years, many individuals come to mind who have contributed to this thesis, and more importantly who have helped me develop as a better researcher and as a better person, overall. I am truly grateful to all these individuals and regret that space permits my mentioning by name only those whose help has been particularly noteworthy.

First and foremost, I'd like to thank my thesis supervisor, Professor Al Oppenheim, for his guidance, encouragement, and support, both moral and financial, especially when they were needed most during the final months of this thesis. I have come to respect him greatly, in part because of his deep interest and involvement in the well-being and professional development of his students, as exemplified by his having established a research environment without equal in the signal processing community. I have benefited immensely from having been a member of his research group and will always be thankful to him for allowing me to be part of that group.

I would also like to thank Professor Udi Weinstein for his constructive feedback on a draft of this thesis and for several enjoyable, stimulating discussions on performance bounds. I found his intensity, insight, and enthusiasm to be truly inspirational. I would also like to thank my two official thesis readers, Professors Al Willsky and Greg Wornell, for their useful advice on the thesis and on potential publications evolving from the thesis.

It goes without saying that I am extremely grateful to my friends and colleagues in the Digital Signal Processing Group (DSPG). I have found working with such an elite group of individuals to be a rewarding, yet often humbling experience. I'd especially like to thank Steve Isabelle and Kambiz Zangi for their friendship and moral support as well as for many fruitful, technical discussions. The three of us struggled to complete our theses at about the same time. Simply knowing that I alone was not undergoing the ordeal of completing a thesis somehow made the completion task less painful for me than it would otherwise have been. Others in the group also deserving special recognition include John Buck, Babis Papadopoulos, Paul Beckmann, Steve Scherock, and Andy Singer, all of whom helped make life at MIT more interesting and enjoyable. I'll never forget my "useful" discussions with them on such diverse topics as dolphin conversations, predicting the price of pork bellies, and methods to avoid being mistaken for a terrorist at a foreign airport. Special thanks goes to our system manager, Giovanni Aliberti, for his friendship and for his resolving my many

computer dilemmas. Special thanks also goes to our group administrative assistant Sally Santiago for her invaluable assistance with many administrative matters involving the MIT bureaucracy as well as for her editorial guidance in my preparation of several conference papers and more importantly this document.

I will always be thankful to my parents for their unceasing encouragement and support and more importantly for their faith and confidence in me, not only while I worked on this thesis but throughout my life. Their inspiring, morale-building phone calls and letters infused warmth and hope in many otherwise dreary, frustrating days. I would also like to thank the rest of my family—my brothers Marc and Ray, my sisters-in-law Lisa and Diane, my nieces Arianne and Kristina, and my nephews David and Kevin—for helping me to maintain a proper perspective on what truly matters in life.

I am also grateful to the Air Force Office of Scientific Research which through its Laboratory Graduate Fellowship Program generously funded four years of this research.

Finally, I'd be remiss if I failed to recognize the contributions of Saint Anthony, Saint Jude, and all the other saints who have worked feverishly on my behalf during these last few years, in response to a deluge of prayers from several of my relatives (and from me). My only regret is that they apparently don't know how to program in C or MATLAB. How appropriate it is that this thesis is being submitted on All Saints Day, the day set aside for honoring saints. I thank these saints and those whose prayers they were responding to.

To my parents, Normand and Elaine, who have have given me so much in life, yet have asked for so little in return, and without whose love, prayers, guidance, and support since the day I was born, I would never have achieved this milestone in my life.

Contents

Chapter 1

Introduction

Chaotic systems have received much attention in the mathematics and physics communities in the last two decades; and they are receiving increasing attention in various engineering disciplines as well. Experimental evidence suggests that these systems may be useful models for a wide variety of physical phenomena, including turbulence, vibrations of buckled elastic structures, and behavior of certain feedback control devices. Traditionally, researchers have focused on possible causes of, or transitions to chaos, universal properties shared by chaotic systems, and various topological and ergodic properties of chaotic systems. As such, the emphasis has been on the analysis of chaotic systems and real-world systems suspected of exhibiting chaos. There has been little attention given to the synthesis of chaotic signals and systems for practical engineering applications such as in communication systems. In part for this reason, useful engineering applications of chaotic systems have yet to appear.

This thesis considers both the analysis and synthesis of chaotic signals and systems. The unifying theme of this thesis is additive noise and the problems and issues involved in dealing with chaotic signals embedded in additive noise. With respect to analysis, the thesis focuses on estimating the state of discrete-time, chaotic systems based on noise-corrupted observations of the state.

The problem of noise-corrupted chaotic signals arises in many applications as does the need for effective state estimation algorithms. For example, often when measurements are taken of a physical phenomenon suspected of being chaotic, the measuring device introduces error in the recorded signal, with the error well-modeled as additive white noise. Alternatively, the actual underlying physical phenomenon may be immersed in a noisy environment,

as might be the case if one seeks to intercept a low-power, chaotic signal, possibly used for secure communication, that has been transmitted over a noisy channel. In both cases, one seeks to separate the chaotic signal from the noise and often to estimate the state of the underlying chaotic system from the noise-corrupted observations of the transformed state. The thesis considers three problem scenarios involving noise-corrupted observations of chaos, with the scenarios distinguished by the level of *a priori* knowledge of the dynamics of the underlying chaotic system. State estimation algorithms are introduced for each of the scenarios, and the performance of the algorithms evaluated using Monte Carlo simulations.

When attempting to design and refine estimators for nonlinear estimation problems, one often has no way of knowing if mediocre performance of an estimator is due to the estimator or to a fundamental aspect of the problem itself. As such, it is often useful and desirable to know the best performance achievable by any estimator for a given nonlinear estimation problem, or equivalently to have upper bounds on achievable state estimator performance. Consequently, state estimator performance bounds are derived and analyzed in the thesis. The analysis reveals that intrinsic properties of chaotic systems—positive Lyapunov exponents and boundedness of attractors—have a fundamental influence on achievable state estimator performance with these systems.

With respect to synthesis, the thesis introduces and analyzes a class of chaotic building blocks having potential, practical value. The basic elements of this class are piecewise linear maps of the unit interval which satisfy certain constraints. These simple maps are shown to exhibit a rich set of properties, which among other things allows computationally efficient, optimal and suboptimal detection as well as maximum-likelihood state estimation with the maps.

The thesis is organized as follows. Chapters 2 and 3 respectively provide background material on chaos and the general state estimation problems considered in this thesis. Specifically, Chapter 2 discusses a number of topological and ergodic properties often associated with chaos, the relations among these properties, as well as the concept of dissipative chaos. Chapter 3 begins by defining the state estimation scenarios of interest in this thesis and continues by briefly reviewing the fundamentals of probabilistic state estimation with an emphasis on the two probabilistic, state estimation approaches focused on the thesis: Maximum-Likelihood (ML) and Minimum-Mean-Squared-Error (MMSE). Next, the optimal MMSE state estimator for linear, dynamical systems—the Kalman filter—is dis-

cussed as well as the practical difficulties in performing optimal MMSE state estimation with nonlinear, dynamical systems. Finally, the chapter provides a summary of previous state estimation research involving nonlinear dynamical systems, in general, followed by a more focused summary of previous state estimation research involving deterministic, chaotic systems.

Chapter 4 introduces state estimation algorithms for discrete-time chaotic systems and time-sampled, continuous-time chaotic systems and provides performance results obtained via Monte Carlo simulations with several chaotic systems. State estimation algorithms are introduced for three problem scenarios—known system dynamics, unknown system dynamics and availability of a noise-free reference orbit, unknown system dynamics and nonavailability of a noise-free reference orbit. In particular, the extended Kalman filter is shown to yield mediocre performance results, but a new variation of this filter is shown to be a potentially effective state estimator for chaotic systems, even when the system dynamics are unknown. A global, approximate MMSE state estimator is also introduced and is shown to be a potentially effective state estimator with input SNRs as small as 0 dB. The estimator exploits intrinsic properties of the steady-state behavior of dissipative, chaotic systems, in particular topological transitivity and ergodicity.

Chapter 5 derives performance bounds for a general class of state estimators and interprets these bounds in the context of chaos. The Cramer-Rao bound for estimators of nonrandom state vectors of deterministic chaotic systems is derived and shown to exhibit behavior similar to that of the corresponding bound for unstable, linear systems. In particular, for dissipative, chaotic diffeomorphisms the bound exhibits a nonzero asymptotic limit when there are noisy observations of the state for only past or only future times. Limitations of the Cramer-Rao bound are discussed, and a specialized form of a general class of performance bounds, the Barankin bounds, is shown to overcome these limitations. The Cramer-Rao and Weiss-Weinstein bounds for estimators of random state vectors are also briefly considered and shown to be of little value for use with chaotic systems.

In contrast to Chapters 4 and 5 which focus on the analysis of chaotic systems and signals produced by these systems, Chapters 6 and 7 focus on the synthesis of chaotic systems. In particular, Chapter 6 introduces a class of piecewise linear maps of the unit interval which give rise to finite state Markov chains. New and previously reported properties of these maps are discussed, and a close relation established between the ergodic properties of these

maps and the Markov chains they give rise to. In addition, these maps are shown to be potentially useful building blocks for maps with arbitrary, invariant probability density functions and for higher-dimensional maps which also give rise to Markov chains. Chapter 7 derives computationally efficient optimal and suboptimal detectors for a subset of these maps and briefly discusses the potential value of these maps and associated detectors for secure communication. The chapter concludes by introducing optimal and suboptimal ML state estimators for use with these maps.

Finally, Chapter 8 provides concluding remarks. The chapter begins by summarizing the highlights of the thesis as well its major contributions to the research community. The chapter then discusses potentially fruitful topics for future research, which build on the results presented in the thesis.

Chapter 2

Overview of Chaos

2.1 Introduction

This chapter establishes a foundation for the state estimation algorithms introduced in Chapter 4 and the performance bounds derived in Chapter 5, by providing a brief introduction to chaos. We discuss a number of properties, both topological and ergodic, often associated with chaos and explain such concepts as invariant measures, attractors, Lyapunov exponents, topological transitivity, and sensitive dependence on initial conditions. We omit discussion of several concepts relevant to chaos but not relevant to this thesis, including dimensions and entropies of attractors. These topics are discussed at varying levels of detail in a number of references on nonlinear dynamical systems, (e.g., [60, 69]).

The discussion of topological and ergodic properties in Section 2.3 is rather formal and abstract, and it uses a number of theoretical concepts from mathematical analysis and topology. Although many of the terms introduced in the section are used throughout the thesis, a thorough understanding of the theory underlying these terms is not needed in order to understand the information on state estimation algorithms and performance bounds provided in Chapters 4 and 5.

2.2 Notational Conventions

We adopt the following notational conventions in this thesis. Plain lowercase and uppercase letters such as x and f denote scalars and scalar-valued functions, whereas bold lowercase and uppercase letters such as $\boldsymbol{x}$ and $\boldsymbol{f}$ denote vectors and vector-valued functions. Except

when confusion might result, functions and arguments of functions are denoted in the text simply by the symbol representing the function or argument. For example, for the scalar equation $y = f(x)$, we say that x is the argument of the function f and y is the value of f evaluated at x, or equivalently the image of x under f.

The character $\mathcal{R}$ denotes the real line and $\mathcal{R}^n$ denotes Euclidean n-space. Unless stated otherwise, all vectors have real-valued components and the domain and range of each function are subsets of $\mathcal{R}^n$ and $\mathcal{R}^m$, respectively, for positive integers m and n. The time index for a time series of scalars or vectors is given in parentheses, so for example $x(n)$ denotes the element of the time series $\{x(i)\}$ at time n. Given two functions g and h, $g \circ h$ denotes the composition of g with h, so that $(g \circ h)(x) = g(h(x))$. Similarly, the shorthand notation $f^n(x) = (\underbrace{f \circ \ldots \circ f}_{n \text{ times}})(x)$ denotes the composition of f with itself n times, and by definition $f^0(x) = x$. We let f^{-1} denote the inverse or inverse image of f, and f^{-n}, the n-fold composition of f^{-1}. For a scalar-valued, differentiable function f, $f'(x)$ denotes the derivative of f evaluated at x, and more generally $f^{(n)}(x)$ denotes the n^{th}-derivative of f evaluated at x. Similarly, for vector-valued, differentiable functions $\boldsymbol{f}$ with vector-valued arguments, $\boldsymbol{D}\{\boldsymbol{f}(\boldsymbol{x})\}$ denotes the derivative of $\boldsymbol{f}$ with respect to $\boldsymbol{x}$, and if the derivative is not taken with respect to the innermost argument, we use a subscript on $\boldsymbol{D}$ to denote the differentiation variable. For example, $\boldsymbol{D}\{\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x}))\}$ denotes the derivative of $\boldsymbol{f} \circ \boldsymbol{g}$ taken with respect to $\boldsymbol{x}$ whereas $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{x})}\{\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x}))\}$ denotes the derivative of $\boldsymbol{f}$ taken with respect to $\boldsymbol{g}(\boldsymbol{x})$.

We let $\log(x)$ denote the natural logarithm of x and $\log_{10}(x)$ denote the logarithm to the base 10 of x. Except when there might be confusion, we use a shorthand notation for probability density functions (PDFs). Specifically, $p(x)$ and $p(y)$ respectively denote the PDFs $p_X(x)$ and $p_Y(y)$ (evaluated at their arguments), and $p(y|x)$ denotes $p_{Y|X}(y|x)$, the PDF of y conditioned on the random variable x. In addition, we use the shorthand notation $\int f d\mu$ to represent the Lebesgue integral of the function f with respect to the measure μ.

The focus in this thesis is on discrete-time, deterministic, dynamical systems, also known as deterministic maps, and time-sampled, continuous-time, deterministic, dynamical systems, also known as deterministic flows. By a deterministic map, we mean an evolution equation specified by a function $\boldsymbol{f} : M \to M$ mapping some space M to itself, which gives rise to a time series $\{\boldsymbol{x}(i)\}$ satisfying the relation $\boldsymbol{x}(n+1) = \boldsymbol{f}(\boldsymbol{x}(n)) = \boldsymbol{f}^n(\boldsymbol{x}(0))$ and if $\boldsymbol{f}$ is invertible the relation $\boldsymbol{x}(0) = \boldsymbol{f}^{-n}(\boldsymbol{x}(n))$. The last two equalities emphasize the fact

that for a deterministic system, the value of the time series at any time uniquely determines the value at all future times and at all past times as well if the system is invertible. We refer to $\boldsymbol{f}$ as a *map*, the time series $\{\boldsymbol{x}(i)\}$ as an *orbit*, and $\boldsymbol{x}(n)$ as either the *state of* $\boldsymbol{f}$ *at time* n or the *orbit point at time* n. To maintain an absolute time reference, we reserve the designation *initial state* or *initial condition* for $\boldsymbol{x}(0)$, the state at time 0, even if our interest is with $\boldsymbol{x}(n)$ for times $n < 0$. We also define an *N-point orbit segment* to be a set of N consecutive points from an orbit, and the *orbit generated by* $\boldsymbol{x}(0)$ to be the unique one-sided orbit $\{\boldsymbol{x}(i)\}_{i=0}^{\infty}$ for noninvertible maps and the unique two-sided orbit $\{\boldsymbol{x}(i)\}_{i=\infty}^{\infty}$ for invertible maps associated with a specific point $\boldsymbol{x}(0)$ at time 0. An orbit O is *periodic* with period n if for each $\boldsymbol{x} \in O$, $(\boldsymbol{f}^n)^i(\boldsymbol{x}) = \boldsymbol{x}$ for all nonnegative integers i and some positive integer n. Note that a periodic orbit has only a finite number of unique points. We refer to points on periodic orbits as *periodic points*. An orbit point $\boldsymbol{x}$ is *asymptotically periodic* if there exists a positive integer N such that the point $\boldsymbol{f}^N(\boldsymbol{x})$ is periodic.

By a deterministic flow, we mean an evolution equation specified by a function $\boldsymbol{F} : M \rightarrow M$ mapping some space M to itself, which gives rise to continuous-time waveforms $\boldsymbol{x}(t)$, $t \in R$ which satisfy the differential equation

$$\frac{d\boldsymbol{x}}{dt} \equiv \boldsymbol{x}'(t) = \boldsymbol{F}(\boldsymbol{x}(t)). \tag{2.1}$$

In this thesis we are primarily interested in the maps which flows give rise to by sampling $\boldsymbol{x}(t)$ every T seconds. That is, the resulting time series $\{\boldsymbol{y}(i)\}$, where $\boldsymbol{y}(n) \equiv \boldsymbol{x}(nT)$, is an orbit of the deterministic map $\boldsymbol{f}_T$ defined as

$$\boldsymbol{x}((n+1)T) = \boldsymbol{f}_T(\boldsymbol{x}(nT)) \equiv \boldsymbol{x}(nT) + \int_{nT}^{(n+1)T} \boldsymbol{F}(\boldsymbol{x}(t))\, dt. \tag{2.2}$$

Note that in contrast to arbitrary maps which may or may not be invertible, the maps which time-sampled flows give rise to are always invertible if $\boldsymbol{F}$ is continous.

2.3 The Distinguishing Properties of Chaos

Deterministic chaotic systems, referred to simply as chaotic systems in the thesis, are nonlinear, deterministic, dynamical systems, either discrete-time or continuous-time, which satisfy a certain set of properties; but, there is no universal agreement as to what these properties

should be. As a result, there is no single definition of deterministic chaos, but instead several, closely related definitions. In this section we discuss topological and ergodic properties often associated with or required for deterministic chaos, discuss the relation among these properties, and cite those properties either satisfied by or believed to be satisfied by the systems considered in this thesis. We only discuss properties relevant to discrete-time and time-sampled continuous-time systems. Topological properties are discussed first, since they are conceptually easier to understand than ergodic properties.

2.3.1 Topological Properties

To discuss the topological properties of chaos, one must first specify a topology [62] on the space M on which the system is defined. For the unit-interval maps considered in Chapters 6 and 7, M is the unit interval and the topology on M is the subspace topology for the metric topology on $\mathcal{R}^n$. In other words, a basis element for the topology is simply the intersection of an open ball in $\mathcal{R}^n$ with M. For the dissipative maps considered in Chapters 4 and 5, M and its associated topology are not as easily defined, since of interest is the steady-state dynamics of such systems and these dynamics typically evolve onto attractors which are not simple subsets of $\mathcal{R}^n$. In the discussion that follows, we implicitly assume that an appropriately defined metric gives rise to the underlying topology.

The three topological properties often required of a dynamical system for it to be considered chaotic are sensitive dependence on initial conditions, topological transitivity, and a dense set of periodic orbits [20]. We briefly consider each in turn.

Most definitions of chaos require that there be *sensitive dependence on initial conditions*, a formal definition of which is the following [20]:

Sensitive Dependence on Initial Conditions: A discrete-time system or map $\boldsymbol{f} : M \rightarrow M$ has sensitive dependence on initial conditions if there exists a constant $\delta > 0$ such that for any $\boldsymbol{x} \in M$ and any neighborhood U of $\boldsymbol{x}$, there exists a $\boldsymbol{y} \in U$ and an integer $n \geq 0$ such that $|\boldsymbol{f}^n(\boldsymbol{x}) - \boldsymbol{f}^n(\boldsymbol{y})| > \delta$.

In other words, for there to be sensitive dependence on initial conditions, there must be a positive constant δ, such that for any point $\boldsymbol{x}$ in M and any arbitrarily small open ball containing $\boldsymbol{x}$, one can always find another point $\boldsymbol{y}$ in that ball such that the distance between corresponding points on the orbits generated by $\boldsymbol{x}$ and $\boldsymbol{y}$ eventually exceeds δ. Intuitively,

this means that there is a *local* separation or divergence of the two orbits. However, the conditions of the definition are still satisfied even if these orbits converge after time n. Also, the definition does not require that the orbits generated by all points in a given neighborhood of $\boldsymbol{x}$ diverge from that generated by $\boldsymbol{x}$. In fact, for many chaotic systems such as the dissipative systems considered in Chapter 4, one can generally find points in each neighborhood of $\boldsymbol{x}$ for which the orbits generated by these points converge to the one generated by $\boldsymbol{x}$.

An unstable, linear system exhibits sensitive dependence on initial conditions, but such a system is not considered chaotic. As such, additional topological properties are generally required of a system for it to be considered chaotic. One such property, not shared by linear systems unless the space M consists only of the origin, is *topological transitivity*, which can be defined as follows [20]:

Topological Transitivity: A map $\boldsymbol{f} : M \rightarrow M$ is topologically transitive if for any pair of open sets $U, V \subset M$, there exists a positive integer k such that $\boldsymbol{f}^k(U) \cap V \neq \emptyset$ where $\emptyset$ is the empty set.

A less abstract definition, applicable when $\boldsymbol{f}$ is continuous and M is compact, is the following [87]:

Topological Transitivity: A continuous transformation $\boldsymbol{f} : M \rightarrow M$ is topologically transitive if there is some $\boldsymbol{x} \in M$ for which the orbit generated by $\boldsymbol{x}$, $O_{\boldsymbol{f}}(x) \equiv \{\boldsymbol{f}^k(\boldsymbol{x}) | k \geq 0\}$, is dense in M. That is, for any $\boldsymbol{y} \in M$ and any open set U containing $\boldsymbol{y}$, there exists a positive integer k such that $\boldsymbol{f}^k(x) \in U$.

Intuitively, this latter definition simply means that the orbit generated by $\boldsymbol{x}$ comes arbitrarily close to every point in M. (More precisely, the above definitions are those for *one-sided* topological transitivity as opposed to *two-sided* topological transitivity, a concept applicable only to invertible systems).

One consequence of topological transitivity is that it prevents M from being decomposable. That is, if $\boldsymbol{f}$ is topologically transitive on M, then one can't divide M into two disjoint subsets M_1 and M_2, such that $\boldsymbol{f}(M_1) \subset M_1$ and $\boldsymbol{f}(M_2) \subset M_2$. As suggested by this indecomposability constraint and discussed in the next section, ergodicity and topological transitivity are closely related. In addition, if $\boldsymbol{f}$ is topologically transitive on M, one can show that if one point has a dense orbit then almost all points in M have dense orbits [87].

In Chapter 4 we exploit this property of chaotic systems to derive simple, yet potentially effective state estimation algorithms.

A third topological property sometimes required of a system for it to be considered chaotic is that there be a dense set of periodic points. Although many if not all of the systems considered in this thesis have this property, it is not a property which we either exploit or emphasize. We do not consider it in the context of state estimation in Chapters 4 and 5 and only briefly consider in the context of signal synthesis in Chapters 6 and 7.

2.3.2 Ergodic Properties

Whereas a discussion of the topological properties associated with chaos requires that a topology be first defined, a discussion of the ergodic properties associated with chaos requires that a measure space (X, β, μ) be defined, where X is a set, β is a σ-algebra of subsets of X, and μ is a measure on β. We only consider probability spaces in this thesis, which are those spaces satisfying $\mu(X) = 1$. One example of a probability space used extensively in Chapters 6 and 7 consists of X being the unit interval, β being the Borel σ-algebra (which by definition is the smallest σ-algebra containing all open subintervals of the unit interval), and μ being Lebesgue measure (the unique measure for which the measure of any subinterval is simply the length of the subinterval). In contrast, for the dissipative maps considered in Chapters 4 and 5, the underlying probability space is rather nebulous, as the steady-state system dynamics occur on a complicated (fractal) attractor, which is singular with respect to Lebesgue measure.

A transformation or map $\boldsymbol{f} : X \rightarrow X$ (i.e., a mapping from X to itself) is *measurable* if $\boldsymbol{f}^{-1}(B) \in \beta$ for every $B \in \beta$, where $\boldsymbol{f}^{-1}$ denotes the inverse image of $\boldsymbol{f}$. A measurable transformation is *nonsingular* if $\mu(\boldsymbol{f}^{-1}(B)) = 0$ for every $B \in \beta$ such that $\mu(B) = 0$. A nonsingular, measurable transformation is *measure-preserving* if $\mu(\boldsymbol{f}^{-1}(B)) = \mu(B)$ for every $B \in \beta$. If in addition to being measure-preserving, the map has the property that the only sets B in β for which $\boldsymbol{f}^{-1}(B) = B$ have measure zero or one, then $\boldsymbol{f}$ is *ergodic* (with respect to the measure μ). In other words, an ergodic map is one for which the only invariant sets are those that contain either all the probability or none of it.

For probability spaces, one consequence of ergodicity is that ensemble averages and infinite time averages correspond, in the sense that the following relation holds for all functions $g \in L^1(\mu)$ (where $L^1(\mu)$ denotes the set of all absolutely integrable, real-valued

functions on (X, β, μ)):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} g(\boldsymbol{f}^i(x)) = \int g \, d\mu \tag{2.3}$$

for almost all x. This relation suggests that ergodicity and topological transitivity are related, and indeed this is the case. As shown in [34], if $\boldsymbol{f}$ is ergodic on the measure space (X, β, μ) and there exists a topology on X with countable basis such that every open set has nonzero measure, then $\boldsymbol{f}$ is topologically transitive with respect to this topology. Thus, ergodicity implies topological transitivity under these conditions. However, an ergodic system need not have sensitive dependence on initial conditions, and thus not all ergodic systems are chaotic.

For the problems considered in this thesis, the map $\boldsymbol{f} : M \to M$ is given and we seek a σ-algebra β on M and a probability measure μ on β such that $\boldsymbol{f}$ is measure-preserving and more importantly ergodic. The measure of interest in Chapters 4 and 5 is the unique, ergodic, so-called *physical measure* [21] which is defined in the following, limiting sense. Consider the noise-driven system

$$\boldsymbol{x}(n+1) = \boldsymbol{f}(\boldsymbol{x}(n)) + \epsilon \boldsymbol{w}(n). \tag{2.4}$$

Under certain conditions this system has a unique invariant measure μ_ϵ, when $\{\boldsymbol{w}(i)\}$ is a white-noise sequence and ϵ is a small positive constant. The physical measure is given by $\lim_{\epsilon\to\infty} \mu_\epsilon$.

The value of an ergodic measure for $\boldsymbol{f}$ arises from the fact that if such a measure μ exists, then by the *Multiplicative Ergodic Theorem of Oseledec* [21], the following limit exists for $\boldsymbol{x}$, μ-almost everywhere (i.e., for all $\boldsymbol{x}$ except on a set of μ-measure zero):

$$\boldsymbol{\Lambda_x} = \lim_{n\to\infty} \left\{ \boldsymbol{D}\{\boldsymbol{f}^n(\boldsymbol{x})^T\} \, \boldsymbol{D}\{\boldsymbol{f}^n(\boldsymbol{x})\} \right\}^{\frac{1}{2n}}, \tag{2.5}$$

and the eigenvalues of $\boldsymbol{\Lambda_x}$ are μ-almost everywhere constant. The logarithms of these eigenvalues are known as the *Lyapunov exponents* of $\boldsymbol{f}$ with respect to the measure μ. In addition, for μ-almost all $\boldsymbol{x} \in \mathcal{R}^{\mathcal{N}}$, if $\lambda_1 > \lambda_2 > \cdots > \lambda_{\mathcal{N}}$ denote the ordered Lyapunov exponents not repeated by multiplicity, and $E^i_{\boldsymbol{x}}$ denotes the subspace of $\mathcal{R}^{\mathcal{N}}$ associated with

all eigenvalues with logarithms that are less than or equal to λ_i, then the following holds:

$$\lim_{n\to\infty} \frac{1}{n} \log \|D\{\boldsymbol{f}^n(\boldsymbol{x})\}\, \boldsymbol{u}\| = \lambda_i \tag{2.6}$$

for each unit vector $\boldsymbol{u} \in E_{\boldsymbol{x}}^i \setminus E_{\boldsymbol{x}}^{i+1}$, where $\|\cdot\|$ denotes the Euclidean norm and $E_{\boldsymbol{x}}^i \setminus E_{\boldsymbol{x}}^{i+1}$ denotes the set formed by taking $E_{\boldsymbol{x}}^i$ and removing the subspace $E_{\boldsymbol{x}}^{i+1}$.

For a linear, time-invariant, deterministic system $\boldsymbol{f}$ with no zero-valued eigenvalues for the matrix representation of $\boldsymbol{f}$, the Lyapunov exponents are the logarithms of the magnitudes of the eigenvalues. For these systems, the multiplicative ergodic theorem implies that the $1/(2n)$ roots of the singular values of $\boldsymbol{f}^n$ converge to the magnitudes of the eigenvalues of $\boldsymbol{f}$, and the subspace spanned by the singular vectors corresponding to the m smallest singular values of $\boldsymbol{f}^n$ converges to the subspace spanned by the m smallest eigenvalues of $\boldsymbol{f}$, for $m = 1, \cdots, \mathcal{N}$.

The relation given by (2.6) suggests that if $\boldsymbol{u} \in E_{\boldsymbol{x}}^i \setminus E_{\boldsymbol{x}}^{i+1}$ is a unit vector and $\boldsymbol{y} = \boldsymbol{x} + \delta \boldsymbol{u}$ where δ is a small positive constant, then

$$\|\boldsymbol{f}^n(\boldsymbol{x}) - \boldsymbol{f}^n(\boldsymbol{y})\| \approx \|D\{\boldsymbol{f}^n(\boldsymbol{x})\}\, \delta \boldsymbol{u}\| \approx \delta \exp(n\lambda_i). \tag{2.7}$$

If the approximation were accurate, an implication would be that $\boldsymbol{f}$ has sensitive dependence on initial conditions if the largest Lyapunov exponent λ_1 is positive. However, this approximation, frequently cited in the literature, is often a poor approximation unless the magnitude of δ is infinitesimally small, and it is thus a poor approximation for practical purposes. As such, it is unclear if the existence of a positive Lyapunov exponent implies that there is sensitive dependence on initial conditions, although at least experimentally this appears to be the case. In addition, since Lyapunov exponents are defined only for ergodic systems, a system with a positive Lyapunov exponent is topologically transitive (with respect to a reasonable topology, e.g., one for which each nonempty open set has nonzero measure).

With most, if not all, definitions of chaos which include ergodic considerations, a fundamental criterion for a system to be considered chaotic is that there be a a positive Lyapunov exponent defined on a nontrivial measure space. The dissipative systems considered in Chapter 4 and 5 satisfy this condition as do a subset of the systems considered in Chapters 6 and 7. Implicitly or explicitly inherent in all these definitions are topological considera-

tions as well, since the definition of Lyapunov exponents assumes the existence of a metric and thus a metric topology on the underlying measure space. As such, whereas one can define chaos solely in terms of topological properties as is done in [20], implicit or explicit in any definition of chaos involving ergodic considerations are topological considerations as well.

2.3.3 Dissipative Chaos

The next two chapters deal with dissipative diffeomorphisms that are either chaotic or suspected of being chaotic. These are perhaps the most interesting and representative of real-world phenomena, yet least understood of all chaotic systems. In particular, the focus is on maps or time-sampled flows, $\boldsymbol{f} : S \rightarrow S$, where $\boldsymbol{f}$ is invertible with both $\boldsymbol{f}$ and $\boldsymbol{f}^{-1}$ having continuous derivatives and where S, the *state space* or *phase space*, is a simply connected (open) subset of $\mathcal{R}^{\mathcal{N}}$ for some $\mathcal{N}$. A *dissipative* map is nonrigorously defined as one which contracts volumes in state space, in contrast to a conservative system which is one that preserves state-space volumes [21].

We briefly discuss dissipative, chaotic maps and their properties in this section. Whereas these maps and their properties remain poorly understand and a rigorous discussion of these properties requires use of mathematical concepts beyond the scope of this thesis, only a brief, nonrigorous discussion is provided here with the reader cautioned that some of the information, much of which was extracted from [21, 33], remains the subject of debate among theoreticians.

One interesting property of dissipative, chaotic systems is that although they are deterministic and contract volumes, their dynamics are nontrivial. In particular, they often give rise to complicated *attractors* known as *strange attractors*. Intuitively, an attractor is a bounded region of phase space, where the points on the orbits generated by initial conditions in some *attracting set* accumulate as n grows large, where an attracting set can be formally defined as follows:

Attracting Set: An attracting set A with *fundamental neighborhood* U is a compact set (in phase space) which is invariant, i.e., $\boldsymbol{f}(A) = A$ and for which every open set V containing A satisfies $\boldsymbol{f}^n(U) \subset V$ for n large enough.

In other words, an attracting set A is one for which orbits generated by points inside A

remain in A for all time, and for which orbits generated by points in a certain open set containing A either eventually enter and remain in A or converge to A.

A slightly more rigorous definition of an attractor is an attracting set which contains a dense orbit, thereby implying that the system $\boldsymbol{f}$ which gives rise to the attractor is topologically transitive when restricted to this attracting set [33, p. 36]. A *strange attractor* can be nonrigorously defined as an attractor for which there is sensitive dependence on initial conditions, with a more rigorous definition being an attractor which contains a transversal homoclinic orbit [33]. An explanation of a transversal homoclinic orbit is beyond the scope of this thesis; its relevance to this discussion is that its existence leads to orbits with nontrivial behavior, with typical orbits often giving rise to complicated, fractal patterns.

The relevance of attractors and strange attractors to this thesis is threefold. First, given an attractor, one can find an open set in $\mathcal{R}^{\mathcal{N}}$ for which the orbit of almost every point in this set converges to the attractor. Second, if there is sensitive dependence on initial conditions (or a transversal homoclinic orbit), the dynamics of typical orbits generated by points on or near the attractor are nontrivial. Third, the steady-state behavior of orbits generated by different points on or near the attractor is similar.

Often a strange attractor has zero volume in the original state space and so-called fractal dimension [21]. A discussion of dimensions of strange attractors is beyond the scope of this thesis. However, as shown in Chapter 4, because of the similar steady-state behavior of orbits generated by points on the attractor and because these orbits occupy a small region of state space, one can derive simple, potentially effective state estimation algorithms for these systems, which do not require full knowledge of the underlying system dynamics.

As noted earlier, a common ergodic criterion for chaos is that there be a positive Lyapunov exponent. However, dissipativeness of a system $\boldsymbol{f}$ generally requires that the Jacobian of $\boldsymbol{f}$ or $\boldsymbol{f}^n$ for some integer n have absolute value less than one. As a (nonobvious) consequence of the multiplicative ergodic theorem and these constraints, a dissipative, chaotic system must have at least one negative Lyapunov exponent, and the sum of the Lyapunov exponents must be negative. Therefore, the state vector dimension must be at least two for a dissipative, chaotic map. Similarly, one can show that a chaotic flow always has a zero-valued Lyapunov exponent which corresponds to motion in the flow direction [21]. Therefore, for a chaotic flow or a map arising from time-sampling the flow, the state vector dimension must be at least three. These conditions on the dimensions of the state vectors

for dissipative, chaotic maps and flows makes analysis of and algorithm development (e.g., state estimation algorithms) for these systems difficult, since one cannot simply develop algorithms for one-dimensional systems and then adapt the results to higher dimensions.

One practical problem that arises in computer simulation of an invertible, dissipative, chaotic map is that the inverse system is generally unstable and the orbits generated by most points rapidly tend to infinity. This follows from the fact that since the system is dissipative and thus contracts volumes, the inverse system expands volumes. Because of this, it is difficult to obtain accurate *backward orbit segments* for points even for those points near the attractor, where a backward orbit segment for a point is an orbit segment for which the point is the final condition.

2.4 Examples of Dissipative, Chaotic Maps and Flows

A number of dissipative maps and flows, which either satisfy or are believed to satisfy fundamental topological and ergodic properties associated with chaos, have been discovered and reported over the last three decades, perhaps the most noteworthy having been Lorenz's seminal discovery of the chaotic flow that bears his name. In this section we discuss the three dissipative systems, two maps and one flow, used for the experimental results presented in Chapters 4 and 5. The three systems are representative of dissipative, chaotic systems, in general, and are perhaps the systems most frequently used in the study of dissipative chaos.

The two dissipative maps used in this thesis are the Henon and Ikeda maps. As with most dissipative systems suspected of being chaotic, the properties of these maps are only partially understood. Both maps are dissipative diffeomorphisms with state vector dimension of two, the minimum dimension for a dissipative, chaotic map. The state or system equations $\boldsymbol{x}(n+1) = \boldsymbol{f}(\boldsymbol{x}(n))$ for the two maps, expressed componentwise are the following:

Henon Map

$$x_1(n+1) = 1 - 1.4\,x_1^2(n) + x_2(n) \tag{2.8}$$

$$x_2(n+1) = .3\,x_1(n) \tag{2.9}$$

Ikeda Map

$$x_1(n+1) = 1 + .9\,[x_1(n)\cos\alpha(n) - x_2(n)\sin\alpha(n)] \tag{2.10}$$

$$x_2(n+1) = .9\,[x_1(n)\,\sin\,\alpha(n) + x_2(n)\,\cos\,\alpha(n)] \tag{2.11}$$

$$\alpha(n) = .4 - \frac{6}{1 + x_1^2(n) + x_2^2(n)}. \tag{2.12}$$

where $\boldsymbol{x}(n) \equiv [x_1(n), x_2(n)]^T$. Other choices for the constants in the above equations have also been used; but, the properties of the resulting maps differ at least slightly from those of the above maps. If the state vector $\boldsymbol{x}(n)$ is treated as a scalar, complex quantity with real and imaginary parts $\boldsymbol{x}_r(n)$ and $\boldsymbol{x}_i(n)$ respectively, so that $\boldsymbol{x}(n) = \boldsymbol{x}_r(n) + j\,\boldsymbol{x}_i(n)$ (where $j^2 = -1$), then the Ikeda map can be succinctly expressed as follows:

$$\boldsymbol{x}(n+1) = 1 + .9\,\boldsymbol{x}(n)\,\exp\left\{j\left[.4 - \frac{6}{1 + \|\boldsymbol{x}(n)\|^2}\right]\right\} \tag{2.13}$$

where $\|\boldsymbol{x}(n)\| = \boldsymbol{x}_r^2(n) + \boldsymbol{x}_i^2(n)$.

Figures 2-1 and 2-2 depict a typical orbit segment for each of the two maps, with a point near the attractor used as the initial condition for each segment. Note that the time ordering of the orbit points in each orbit segment is not discernible from the figures. The

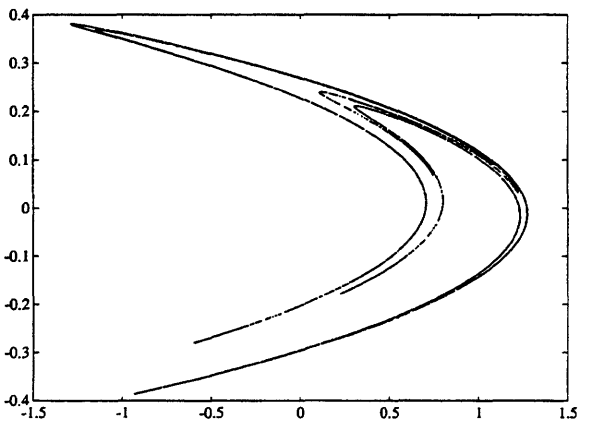

Figure 2-1: Orbit Segment for the Henon Map

orbit segment for each map traces out the complicated attractor associated with that map. Because of the ergodic nature of the maps, orbit segments generated by most other initial conditions near the attractor trace out the same patterns.

The chaotic flow used in this thesis is the Lorenz flow, perhaps the most widely investigated of all chaotic systems. The state dimension for this flow is three, the minimum dimension for a dissipative, chaotic flow. The state or system equations $\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{F}(\boldsymbol{x}(t))$

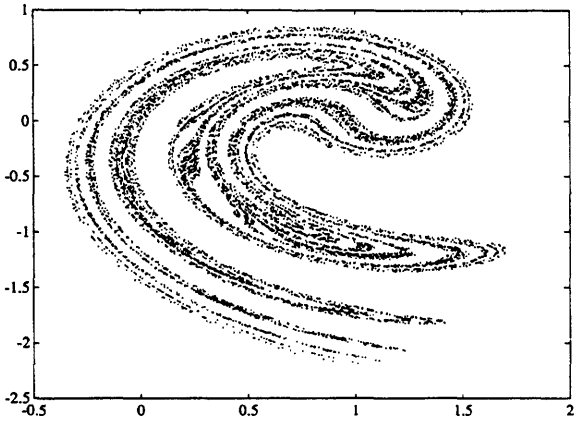

Figure 2-2: Orbit Segment for the Ikeda Map

for the flow, expressed componentwise, are given by

$$\frac{dx_1(t)}{dt} = 10[x_2(t) - x_1(t)] \tag{2.14}$$

$$\frac{dx_2(t)}{dt} = 28\,x_1(t) - x_2(t) - x_1(t)\,x_3(t) \tag{2.15}$$

$$\frac{dx_3(t)}{dt} = -\frac{8}{3}\,x_3(t) + x_1(t)\,x_2(t). \tag{2.16}$$

where $\boldsymbol{x}(t) = [x_1(t), x_2(t), x_3(t)]^T$. Figure 2-3 depicts a typical trajectory projected onto the (x_1, x_3) plane for the Lorenz flow. The trajectory shown in the figure and the Lorenz

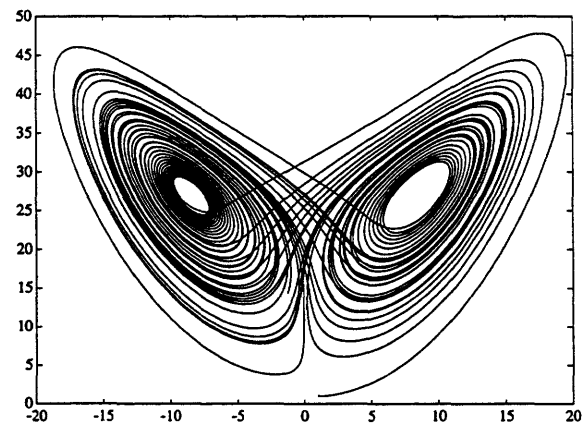

Figure 2-3: Projection of Lorenz Trajectory onto (x_1, x_3) Plane

trajectories used throughout the thesis were obtained by numerically integrating the state

equation using the fourth-order Runge Kutta method with a step size of .005.

Chapter 3

State Estimation Fundamentals

3.1 Introduction

In this chapter we establish a foundation for the estimation problems, algorithms, and performance bounds presented in Chapters 4 and 5. The chapter begins by introducing the state estimation scenario of interest in this thesis and then briefly reviews the two probabilistic estimation techniques—Maximum-Likelihood (ML) and Minimum-Mean-Squared-Error (MMSE)—which underlie the state estimation algorithms discussed in Chapter 4. The chapter continues by briefly discussing the Kalman filter, the optimal, MMSE state estimator for linear, state estimation problems. Next, the chapter provides a historical summary of nonlinear, state estimation research in general and concludes with a more focused summary of state estimation research involving chaotic systems.

3.2 State Estimation Scenario

Of the three general, nonlinear, state estimation scenarios traditionally considered in the estimation and control literature, two are relevant to this thesis. The first, referred to as the DTS/DTO scenario in the thesis, consists of discrete-time state and observation equations with commonly used but not the most general forms of the equations given by the following:

Discrete-Time System, Discrete-Time Observation (DTS/DTO)

$$\boldsymbol{x}(n+1) = \boldsymbol{f}_n(\boldsymbol{x}(n)) + \boldsymbol{g}_n(\boldsymbol{x}(n))\,\boldsymbol{w}(n) \tag{3.1}$$

$$\boldsymbol{y}(n) = \boldsymbol{h}_n(\boldsymbol{x}(n)) + \boldsymbol{v}(n). \tag{3.2}$$

In the first equation, the *state equation*, $\boldsymbol{x}(n)$ is the $\mathcal{N}$-dimensional *state vector* we seek to estimate; $\boldsymbol{f}_n$ and $\boldsymbol{g}_n$ are nonlinear, possibly time-varying functions of the state which are usually required to satisfy certain smoothness constraints; and $\boldsymbol{w}(n)$, the *driving noise*, is an $\mathcal{N}$-dimensional, zero-mean, Gaussian, white-noise process. In the second equation, the *observation equation*, $\boldsymbol{y}(n)$ is the $\mathcal{P}$-dimensional *observation vector* used for estimating $\boldsymbol{x}(n)$; $\boldsymbol{h}_n$ is a nonlinear, possibly time-varying function of the state which is usually required to satisfy certain smoothness constraints; and $\boldsymbol{v}(n)$, the *observation noise*, is a $\mathcal{P}$-dimensional, zero-mean, Gaussian, white-noise process. Generally, $\boldsymbol{w}(n)$ and $\boldsymbol{v}(n)$ are assumed to be uncorrelated with each other and with the initial condition $\boldsymbol{x}(0)$.

The second scenario, referred to as the CTS/DTO scenario in the thesis, consists of a continuous-time state equation and a discrete-time observation equation with commonly used, but not the most general, forms of the equations given by the following:

Continuous-Time System, Discrete-Time Observation (CTS/DTO)

$$d\boldsymbol{x}(t) = \boldsymbol{F}_t(\boldsymbol{x}(t))\,dt + \boldsymbol{G}_t(\boldsymbol{x}(t))\,d\boldsymbol{W}(t) \tag{3.3}$$

$$\boldsymbol{y}(n) = \boldsymbol{h}_n(\boldsymbol{x}(nT)) + \boldsymbol{v}(n). \tag{3.4}$$

For this scenario, the state equation is a stochastic differential equation [42, 54] in which $\boldsymbol{x}(t)$ is the $\mathcal{N}$-dimensional state vector we seek to estimate; $\boldsymbol{F}_t$ and $\boldsymbol{G}_t$ are nonlinear, possibly time-varying functions of the state which are required to satisfy a set of both smoothness and growth-rate constraints; and $\boldsymbol{W}(t)$ is an $\mathcal{N}$-dimensional standard Brownian motion. The observation equation for this scenario has the same interpretation as for the DTS/DTO scenario.

Whereas the focus of this thesis is on chaos, only restricted forms of the DTS/DTO and CTS/DTO scenarios are of interest. In particular, we require the functions $\boldsymbol{f}_n$, $\boldsymbol{F}_t$, $\boldsymbol{g}_n$, $\boldsymbol{G}_t$ and $\boldsymbol{h}_n$ in (3.1–3.4) to be time-invariant and thus expressible as $\boldsymbol{f}$, $\boldsymbol{F}$, $\boldsymbol{g}$, $\boldsymbol{G}$ and $\boldsymbol{h}$. In addition, we require $\boldsymbol{f}$ in (3.1) to be a chaotic map and $\boldsymbol{F}$ in (3.3) to be a chaotic flow; and in Chapters 4 and 5 we further require that $\boldsymbol{f}$ and $\boldsymbol{F}$ be dissipative diffeomorphisms. Finally, whereas our interest is in *deterministic* systems and the properties exhibited by a class of these systems, we consider the restricted form of the state equations in which driving noise is absent. With these restrictions, the equations for the DTS/DTO model

reduce to the following:

$$\boldsymbol{x}(n+1) = \boldsymbol{f}(\boldsymbol{x}(n)) \tag{3.5}$$

$$\boldsymbol{y}(n) = \boldsymbol{h}(\boldsymbol{x}(n)) + \boldsymbol{v}(n) \tag{3.6}$$

and the equations for the CTS/DTO scenario reduce to the following:

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{F}(\boldsymbol{x}(t)) \tag{3.7}$$

$$\boldsymbol{y}(n) = \boldsymbol{h}(\boldsymbol{x}(nT)) + \boldsymbol{v}(n) \tag{3.8}$$

where $\boldsymbol{f}$ is a dissipative, chaotic diffeomorphism and $\boldsymbol{F}$ is a dissipative, chaotic flow.

The omission of driving noise in (3.5) and (3.7) renders these state equations fundamentally different from the more general equations (3.1) and (3.3) respectively, from which they came. In particular, the stochastic nature of the processes $\boldsymbol{x}(n)$ and $\boldsymbol{x}(t)$ which (3.1) and (3.3) respectively give rise to is due both to uncertainty in the initial condition $\boldsymbol{x}(0)$ and to the driving noise terms $\boldsymbol{w}(n)$ and $\boldsymbol{W}(t)$. In contrast, the stochastic nature of the processes which (3.5) and (3.7) give rise to is due solely to uncertainty in the initial condition. That is, if the initial condition is known with certainty, the state at all future times, and at all past times if the system is invertible, is known with certainty as well regardless of the observation noise $\boldsymbol{v}(n)$. Consequently, the deterministic problem considered in this thesis is a simpler problem than the one involving a noise-driven state equation and facilitates the derivation of potentially effective, albeit heuristic, state estimation algorithms. In addition, the derivation and interpretation of performance bounds for the deterministic problem, as is done in Chapter 5, is a far simpler task than the derivation and interpretation of bounds for the problem involving a noise-driven state equation.

Nonetheless, state estimation involving a deterministic, chaotic system has many similarities to state estimation involving a noise-driven system, as one deals with nontrivial, invariant measures and positive entropy rates for both systems. In addition, because chaotic systems exhibit sensitive dependence on initial conditions and because round-off error is inevitable in computer simulations involving chaos, a state equation with a small driving noise term is sometimes a more representative model of the underlying system dynamics when computer-generated chaos is being dealt with. Although we do not adopt such a

model in this thesis, in the next chapter we address the problems that arise from computer round-off error when designing practical, state estimation algorithms and offer simple, albeit suboptimal remedies for these problems.

As suggested in Chapter 2, we can express the restricted CTS/DTO scenario given by (3.7) and (3.8) as a DTS/DTO scenario. In particular, since it is deterministic, the time-sampled state equation (3.7) is given by the following discrete-time state equation:

$$\begin{aligned} \boldsymbol{z}(n+1) &= \boldsymbol{f}_T(\boldsymbol{z}(n)) && (3.9) \\ \text{where } \boldsymbol{z}(n) &\equiv \boldsymbol{x}(nT) && (3.10) \\ \boldsymbol{f}_T(\boldsymbol{z}(n)) &\equiv \boldsymbol{z}(n) + \int_{nT}^{(n+1)T} \boldsymbol{F}(\boldsymbol{x}(t))\, dt. && (3.11) \end{aligned}$$

However, as mentioned in Chapter 2, chaotic maps arising from time-sampled chaotic flows have certain properties not shared by all chaotic maps. In particular, since a differentiable flow is always invertible, the same holds for any map that arises by time-sampling the flow. Also, as discussed earlier, there is a minimum state vector dimension of three for a chaotic flow, and the flow must have at least one zero-valued Lyapunov exponent. The same applies to maps that arise by time-sampling the flow.

3.3 Maximum-Likelihood (ML) and Minimum-Mean-Squared-Error (MMSE) State Estimation

The focus in this thesis is on estimating the state $\boldsymbol{x}(n_0)$ at either a fixed time n_0 or a sequential set of times for the restricted DTS/DTO and CTS/DTO problem scenarios introduced in the previous section, using a given set of observations $Y = \{\boldsymbol{y}(i)\}$ which generally will be sequential in time and thus expressible as $Y(M,N) \equiv \{\boldsymbol{y}(i)\}_{i=M}^{N}$ where M and N are integers with $M < N$. Both filtering and smoothing are considered, where filtering involves estimating $\boldsymbol{x}(n_0)$, the state at time n_0, using observations $\boldsymbol{y}(i)$ only for times $i \leq n_0$, whereas smoothing involves estimating $\boldsymbol{x}(n_0)$ using observations for times $i > n_0$ as well.

Two probabilistic, state estimation approaches that have proven useful in many applications [52, 72, 86] are emphasized—Maximum-Likelihood (ML) and Minimum-Mean-Squared-Error (MMSE). Recall that with ML parameter estimation, the unknown parameter one seeks to estimate is treated as a nonrandom quantity and the ML estimate is

that value of the parameter which maximizes an appropriately defined *likelihood function* or equivalently the logarithm of the likelihood function. For the problem of interest here, the unknown, nonrandom parameter is $\boldsymbol{x}(n_0)$, a set of observations Y is given, and the ML estimate of $\boldsymbol{x}(n_0)$, hereafter denoted $\hat{\boldsymbol{x}}_{ML}(n_0)$, is that value of $\boldsymbol{x}(n_0)$ which maximizes the likelihood function $p(Y;\boldsymbol{x}(n_0))$, where $p(Y;\boldsymbol{x}(n_0))$ denotes the probability density function (PDF) of the observation set Y for a given $\boldsymbol{x}(n_0)$, with an underlying assumption being that the PDF exists. For the restricted DTS/DTO scenario, it follows from (3.6) and the assumptions on the observation noise sequence $\{\boldsymbol{v}(n)\}$ that $\log p(\boldsymbol{y}(i);\boldsymbol{x}(i))$ is given by

$$\begin{aligned}\log p(\boldsymbol{y}(i);\boldsymbol{x}(i)) &= \log(2\pi|\boldsymbol{R}|)^{\frac{-\mathcal{P}}{2}} \\ &\quad -\frac{1}{2}[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{x}(i))]^T\boldsymbol{R}^{-1}[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{x}(i))]\end{aligned} \tag{3.12}$$

where $\mathcal{P}$ is the dimension of the observation vector and $\boldsymbol{R}$ is the covariance matrix of $\boldsymbol{v}(n)$. In light of the determinism of the state equation (3.5) and the assumed invertibility of $\boldsymbol{f}$, we also have

$$\begin{aligned}\log p(\boldsymbol{y}(i);\boldsymbol{x}(n_0)) &= \log(2\pi|\boldsymbol{R}|)^{\frac{-\mathcal{P}}{2}} \\ &\quad -\frac{1}{2}\left[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right]^T\boldsymbol{R}^{-1}\left[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right].\end{aligned} \tag{3.13}$$

Using this equality and exploiting the whiteness of the observation noise leads to the following expression for the log-likelihood function $\log p(Y(M,N);\boldsymbol{x}(n_0))$ for the restricted DTS/DTO scenario:

$$\begin{aligned}\log p(Y(M,N);\boldsymbol{x}(n_0)) &= \log(2\pi|\boldsymbol{R}|)^{\frac{-(N-M+1)\mathcal{P}}{2}} \\ &\quad -\frac{1}{2}\sum_{i=M}^{N}\left[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right]^T\boldsymbol{R}^{-1}\left[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right].\end{aligned} \tag{3.14}$$

The log-likelihood function for the CTS/DTO scenario has a similar form.

Also recall that with MMSE estimation, in contrast to ML estimation, the unknown parameter one seeks to estimate is treated as a random quantity. For the problem of interest here with $\boldsymbol{x}(n_0)$ the unknown parameter to be estimated and Y a given observation

set, the mean-squared estimation error is given by

$$\int\int \|\hat{\boldsymbol{x}}(n_0) - \boldsymbol{x}(n_0)\|^2 \, p(\boldsymbol{x}(n_0), Y) \, d\boldsymbol{x}(n_0) \, dY \tag{3.15}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\hat{\boldsymbol{x}}(n_0)$ denotes an arbitrary estimator for $\boldsymbol{x}(n_0)$, $p(\boldsymbol{x}(n_0), Y)$ denotes the joint PDF of the state vector $\boldsymbol{x}(n_0)$ and observation set Y, and where the integration is over the state vector $\boldsymbol{x}(n_0)$ and the entire observation set Y. A fundamental result in estimation theory is that the MMSE estimator results by choosing the conditional mean $E(\boldsymbol{x}(n_0)|Y)$ as the estimate of $\boldsymbol{x}(n_0)$ for each observation set Y, where $E(\boldsymbol{x}(n_0)|Y)$ is given by

$$E(\boldsymbol{x}(n_0)|Y) = \int \boldsymbol{x}(n_0) \, p(\boldsymbol{x}(n_0)|Y) \, d\boldsymbol{x}(n_0), \tag{3.16}$$

and where $p(\boldsymbol{x}(n_0)|Y)$, the *a posteriori* state density, is the PDF of $\boldsymbol{x}(n_0)$ conditioned on the observation set Y. Use of Bayes rule allows $E(\boldsymbol{x}(n_0)|Y)$ also to be expressed as

$$E(\boldsymbol{x}(n_0)|Y) = \frac{\int \boldsymbol{x}(n_0) \, p(Y|\boldsymbol{x}(n_0)) \, p(\boldsymbol{x}(n_0)) \, d\boldsymbol{x}(n_0)}{\int p(Y|\boldsymbol{x}(n_0)) \, p(\boldsymbol{x}(n_0)) \, d\boldsymbol{x}(n_0)}, \tag{3.17}$$

where $p(\boldsymbol{x}(n_0))$ denotes the unconditional or *a priori* PDF of $\boldsymbol{x}(n_0)$ and $p(Y|\boldsymbol{x}(n_0))$ denotes the PDF of Y conditioned on $\boldsymbol{x}(n_0)$. Note that $p(Y|\boldsymbol{x}(n_0))$ has the same form as the PDF $p(Y; \boldsymbol{x}(n_0))$ defined earlier, with the difference between the two PDFs being that $\boldsymbol{x}(n_0)$ is a random vector in the former and a nonrandom vector in the latter.

An inherent assumption in (3.15) is the existence of the joint PDF $p(\boldsymbol{x}(n_0), Y)$ with respect to the product measure $d\boldsymbol{x}(n_0) \, dY$; an inherent assumption in (3.16) is the existence of the conditional PDF $p(\boldsymbol{x}(n_0)|Y)$ with respect to the measure $d\boldsymbol{x}(n_0)$ (i.e., Lebesgue measure on $\mathcal{R}^{\mathcal{N}}$ where $\mathcal{N}$ is the dimension of $\boldsymbol{x}(n_0)$); and an inherent assumption in (3.17) is the existence of the PDF $p(\boldsymbol{x}(n_0))$ with respect to the measure $d\boldsymbol{x}(n_0)$ and the conditional PDF $p(Y|\boldsymbol{x}(n_0))$ with respect to the measure dY (i.e., Lebesgue measure on $\mathcal{R}^{(N-M+1)\mathcal{P}}$ where $N - M + 1$ is the number of observations and $\mathcal{P}$ is the dimension of each observation vector). For dissipative, chaotic systems, these assumptions are not necessarily valid. For example, if the only *a priori* knowledge about $\boldsymbol{x}(n_0)$ is that it lies on the attractor, then an appropriate *a priori* distribution for $\boldsymbol{x}(n_0)$ is given by the physical measure on the attractor. Intuitively, this *a priori* distribution corresponds to $\boldsymbol{x}(n_0)$ having the same likelihood of

being at each point on an (infinitely long) orbit on the attractor. However, as mentioned earlier, for dissipative, chaotic systems the physical measure is often singular with respect to Lebesgue measure on $\mathcal{R}^{\mathcal{N}}$ and thus has no corresponding PDF with respect to Lebesgue measure. Consequently, the PDFs $p(\boldsymbol{x}(n_0))$, $p(\boldsymbol{x}(n_0), Y)$, and $p(\boldsymbol{x}(n_0)|Y)$ are generally not defined for dissipative systems if the *a priori* distribution of $\boldsymbol{x}(n_0)$ is given by the physical measure on the attractor. In contrast, the likelihood function $p(Y|\boldsymbol{x}(n_0))$ is well-defined in these situations because of our assumptions on the observation noise.

When the joint density $p(\boldsymbol{x}(n_0), Y)$ is not defined, we can express the MSE with respect to the joint probability measure on $\boldsymbol{x}(n_0)$ and Y, or alternatively as follows:

$$\int \int \|\hat{\boldsymbol{x}}(n_0) - \boldsymbol{x}(n_0)\|^2 \, p(Y|\boldsymbol{x}(n_0)) \, d\mu_{\boldsymbol{x}(n_0)} \, dY \tag{3.18}$$

where $\mu_{\boldsymbol{x}(n_0)}$ denotes the *a priori* distribution of $\boldsymbol{x}(n_0)$, and the integration over $\boldsymbol{x}(n_0)$ is defined as a Lebesgue integral. Similarly, we can express the conditional mean as

$$E(\boldsymbol{x}(n_0)|Y) = \frac{\int \boldsymbol{x}(n_0) \, p(Y|\boldsymbol{x}(n_0)) \, d\mu_{\boldsymbol{x}(n_0)}}{\int p(Y|\boldsymbol{x}(n_0)) \, d\mu_{\boldsymbol{x}(n_0)}}. \tag{3.19}$$

As we show in Chapter 4, although the above definition for the conditional mean is abstract and not computable in practice, it can be cast in a revealing form that gives rise to a converging sequence of simple, approximate MMSE state estimators.

3.4 Linear State Estimation and the Kalman Filter

The Kalman filter is a computationally efficient, recursive MMSE state estimator for both continuous-time and discrete-time, linear, state-space models with certain restrictions on the driving noise, observation noise, and distribution of the initial state [5, 30, 38, 57]. Of relevance to this thesis is the form of the Kalman filter applicable to the following, linear, DTS/DTO scenario:

$$\boldsymbol{x}(n+1) = \boldsymbol{F}_n \boldsymbol{x}(n) + \boldsymbol{G}_n \boldsymbol{w}(n) \tag{3.20}$$

$$\boldsymbol{y}(n) = \boldsymbol{H}_n \boldsymbol{x}(n) + \boldsymbol{v}(n), \tag{3.21}$$

where $\boldsymbol{F}_n$, $\boldsymbol{G}_n$, and $\boldsymbol{H}_n$ are matrices, $\boldsymbol{w}(n) \sim \Phi(\boldsymbol{0}, \boldsymbol{Q}_n)$, $\boldsymbol{v}(n) \sim \Phi(\boldsymbol{0}, \boldsymbol{R}_n)$, $\boldsymbol{x}(0) \sim \Phi(\boldsymbol{m}_0, \boldsymbol{P}_0)$ (where $\Phi(\boldsymbol{m}, \boldsymbol{P})$ denotes the normal distribution with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{P}$), and where the driving noise $\boldsymbol{w}(n)$ and observation noise $\boldsymbol{v}(n)$ are independent of each other and the initial state.

The Kalman filter exploits the fact that for linear state estimation, the *a posteriori* density $p(\boldsymbol{x}(n)|Y(0,n))$ is Gaussian and that the mean of the density is the MMSE state estimate. The Kalman filter uses a two-step procedure for recursively computing two quantities—the state estimate $\hat{\boldsymbol{x}}(n)$ and the error covariance matrix $\boldsymbol{P}(n)$ given by

$$\boldsymbol{P}(n) \equiv E\left[(\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n))(\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n))^T)|Y(0,n)\right]. \tag{3.22}$$

In the first step, the *prediction step*, the state estimate and covariance matrix for time $n+1$ are computed from the final state estimate and covariance matrix for time n. In the second step, the *measurement (*or *observation) update* step, the quantities calculated in the first step are updated using the new observation $\boldsymbol{y}(n+1)$. As is the usual convention, we let $\hat{\boldsymbol{x}}(n+1|n)$ and $\boldsymbol{P}(n+1|n)$ denote the state estimate and error covariance matrix, respectively, computed in the prediction step, with the notation chosen to emphasize the fact that these quantities are for time $n+1$ based on observations through time n. Similarly, we let $\hat{\boldsymbol{x}}(n+1|n+1)$ and $\boldsymbol{P}(n+1|n+1)$ denote the updated estimates calculated in the measurement update step, with the notation chosen to emphasize the fact that these quantities are for time $n+1$ based on observations through time $n+1$. These definitions are used in Table 3.1, which provides the equations for the two-step estimation procedure constituting the DTS/DTO Kalman filter.

For many applications, there is an improvement in state estimation performance over filtering if smoothing is used. As shown in the next chapter, this is especially true with chaotic systems. Historically, research has focused on three classes of smoothing problems. The first, *fixed-point* smoothing, involves estimating the state vector $\boldsymbol{x}(n)$ based on the observation set $Y(0,m) = \{\boldsymbol{y}(i)\}_{i=0}^{m}$ for a fixed time n and increasing m, where $m > n$. The second, *fixed-lag* smoothing, involves estimating the state vector $\boldsymbol{x}(n-L)$ based on the observation set $Y(0,n) = \{\boldsymbol{y}(i)\}_{i=0}^{n}$ for each time n and a fixed lag L. The third, *fixed-interval* smoothing, involves estimating the state vector $\boldsymbol{x}(n)$ based on the observation set $Y(0,N) = \{\boldsymbol{y}(i)\}_{i=0}^{N}$ for all times n satisfying $0 \leq n \leq N$.

Prediction Step

$$\hat{\boldsymbol{x}}(n+1|n) = \boldsymbol{F}_n\hat{\boldsymbol{x}}(n|n) \quad (3.23)$$

$$\boldsymbol{P}(n+1|n) = \boldsymbol{F}_n\boldsymbol{P}(n|n)\boldsymbol{F}_n^T + \boldsymbol{G}_n\boldsymbol{Q}_n\boldsymbol{G}_n^T \quad (3.24)$$

Measurement Update Step

$$\hat{\boldsymbol{x}}(n+1|n+1) = \hat{\boldsymbol{x}}(n+1|n) + \boldsymbol{K}(n+1)[\,\boldsymbol{y}(n+1) - \boldsymbol{H}_{n+1}\hat{\boldsymbol{x}}(n+1|n)] \quad (3.25)$$

$$\boldsymbol{K}(n+1) = \boldsymbol{P}(n+1|n)\,\boldsymbol{H}_{n+1}^T\left[\boldsymbol{H}_{n+1}\,\boldsymbol{P}(n+1|n)\,\boldsymbol{H}_{n+1}^T + \boldsymbol{R}(n+1)\right]^{-1} \quad (3.26)$$

$$\boldsymbol{P}(n+1|n+1) = [\boldsymbol{I}_{\mathcal{N}} - \boldsymbol{K}(n+1)\boldsymbol{H}_{n+1}]\,\boldsymbol{P}(n+1|n) \quad (3.27)$$

Initialization

$$\hat{\boldsymbol{x}}(0|-1) = \boldsymbol{m}_0 \quad (3.28)$$

$$\boldsymbol{P}(0|-1) = \boldsymbol{P}_0. \quad (3.29)$$

Table 3.1: The Kalman Filter Equations for the DTS/DTO Model

In an earlier report [76], we considered fixed-interval smoothing in the context of chaos. In this thesis, we consider fixed-lag smoothing which we have found to offer comparable if not superior performance results with chaotic systems. There are various methods, all equivalent, for recursively computing the fixed lag estimate $\hat{\boldsymbol{x}}(n-L|n)$. The most straightforward [5] involves first forming an extended state vector $\boldsymbol{X}(n) \equiv [\boldsymbol{x}(n)^T, \boldsymbol{x}(n-1)^T, \cdots, \boldsymbol{x}(n-L)^T]^T$, applying the Kalman filter equations to the system obtained with this state vector, and noting that various submatrices of the error covariance matrix for the extended system can be updated separately with only some of these needed for recursively estimating $\boldsymbol{X}(n)$. With this approach, one not only obtains the desired, fixed-lag estimate $\hat{\boldsymbol{x}}(n-L|n)$, but the estimates $\hat{\boldsymbol{x}}(n-i|n)$ for $i = 1, 2, \cdots L-1$. The resulting fixed-lag smoothing equations for the estimates $\hat{\boldsymbol{x}}(n-i|n)$ for $i = 1, 2, \cdots L$ are provided in Table 3.2. These equations supplement the Kalman filtering equations provided in Table 3.1

3.5 Nonlinear State Estimation

There has been much research in the past on nonlinear state estimation, most notably a flurry of activity in the early 1970's involving both discrete-time and continuous-time systems and a second wave of activity focusing on continuous-time systems in the early 1980's. Until recently, nearly all the research has dealt with noise-driven state equations

Definitions

$$\hat{\boldsymbol{x}}_i(n|n) \equiv \hat{\boldsymbol{x}}(n-i|n) \;\; i = 1, \cdots, L \tag{3.30}$$

$$\hat{\boldsymbol{x}}_0(n|n) \equiv \hat{\boldsymbol{x}}(n|n) \text{ (from Kalman filter equations)} \tag{3.31}$$

$$\boldsymbol{P}_{0,0}(n|n) \equiv \boldsymbol{P}(n|n) \text{ (from Kalman filter equations)} \tag{3.32}$$

Prediction Step

$$\hat{\boldsymbol{x}}_i(n+1|n) = \hat{\boldsymbol{x}}_{i-1}(n|n) \tag{3.33}$$

$$\boldsymbol{P}_{i,i}(n+1|n) = \boldsymbol{P}_{i-1,i-1}(n|n) \tag{3.34}$$

$$\boldsymbol{P}_{i,0}(n+1|n) = \boldsymbol{P}_{i-1,0}(n|n)\boldsymbol{F}_n^T \tag{3.35}$$

Measurement Update Step

$$\hat{\boldsymbol{x}}_i(n+1|n+1) = \hat{\boldsymbol{x}}_i(n|n) + \boldsymbol{K}_i(n+1)[\boldsymbol{y}(n+1) - \boldsymbol{H}_{n+1}\hat{\boldsymbol{x}}(n+1|n)] \tag{3.36}$$

$$\begin{aligned}\boldsymbol{K}_i(n+1) = {} & \boldsymbol{P}_{i,0}(n+1|n)\,\boldsymbol{H}_{n+1}^T \left[\boldsymbol{H}_{n+1}\,\boldsymbol{P}_{0,0}(n+1|n)\,\boldsymbol{H}_{n+1}^T + \boldsymbol{R}(n+1)\right]^{-1} \\ & \times \boldsymbol{H}_{n+1}\,\boldsymbol{P}_{i,0}^T(n+1|n)\end{aligned} \tag{3.37}$$

$$\boldsymbol{P}_{i,i}(n+1|n+1) = \boldsymbol{P}_{i,i}(n+1|n) - \boldsymbol{K}_i(n+1) \tag{3.38}$$

$$\boldsymbol{P}_{i,0}(n+1|n+1) = \boldsymbol{P}_{i,0}(n+1|n) - \boldsymbol{K}_i(n+1) \tag{3.39}$$

Initialization

$$\hat{\boldsymbol{x}}_i(0|-1) = \boldsymbol{0} \tag{3.40}$$

$$\boldsymbol{P}_{i,i}(0|-1) = [\boldsymbol{0}] \tag{3.41}$$

$$\boldsymbol{P}_{i,0}(0|-1) = [\boldsymbol{0}] \tag{3.42}$$

Table 3.2: The Fixed-Lag Smoothing Equations for the DTS/DTO Model

with nonchaotic systems. Also, the emphasis has been on the classical, recursive, filtering problem in which all parameters in the system model are known, and the objective is to estimate the state at time n using only the estimate at time $n-1$ and the observation at time n. With few exceptions, the focus has been on recursive computation of the *a posteriori* PDF $p(\boldsymbol{x}(n_0)|Y)$ used for MMSE state estimation.

The fundamental problem encountered with most nonlinear systems is that this density requires an infinite number of parameters (e.g., moments) for its specification at each time instant. As a result, optimal, recursive, MMSE state estimation with these systems entails propagating an infinite set of parameters through time. Although it is straightforward to derive a recursive equation, known as the forward-Kolmogorov or Wiener-Hopf equation, for updating the *a posteriori* PDF, simplifying approximations are needed for implementing the equation in practical applications. Consequently, the challenge for nonlinear, state estimation researchers has been and remains the development of practical algorithms for approximating and recursively updating the *a posteriori* PDF, ideally in some optimal sense.

The situation of an infinite dimensional *a posteriori* PDF, encountered in many nonlinear, state estimation problems, contrasts markedly with that for linear, state estimation problems. As discussed in the previous section, under certain conditions the *a posteriori* PDF for linear, state estimation problems is Gaussian and thus completely characterized by two finite sets of parameters—a mean vector and a covariance matrix—with the mean vector being the MMSE state estimate.

Historically, two broad classes of techniques for nonlinear state estimation have been pursued: local and global. Local techniques share the property of using a single, time-varying reference point to calculate quantities, e.g., derivatives, moments, or series coefficients at each time instant. These techniques generally work best when the *a posteriori* PDF is nearly Gaussian or at least unimodal. Included among these techniques are those which approximate and propagate a finite set of low-order moments of the *a posteriori* PDF, as well as series expansion techniques in which the *a posteriori* PDF is represented by a series expansion and a finite number of coefficients propagated through time [57, 83, 84].

The most popular local technique and indeed the most popular of all techniques for nonlinear state estimation has been the extended Kalman filter (EKF). Its popularity is primarily due to its simplicity, ease of implementation, and surprisingly good performance with many nonlinear systems. In the next chapter, we show that the EKF can be incorpo-

rated in a state estimation algorithm for chaotic systems, which is potentially effective even when the system dynamics are unknown.

The unifying element of global techniques is the numerical evaluation of the *a posteriori* PDF with the use of a possibly time-varying, finite grid of points or regions at which calculations and approximations are made [83, 84]. Many global techniques emerged during the early 1970's. Among the most popular was the Gaussian sum approach of Allspach and Sorenson [4], which essentially consisted of a combination of extended Kalman filters operating in parallel. In addition, Bucy and Senne [15] developed and refined a point-mass approach which entailed approximating the *a posteriori* density with a finite set of impulses or point masses defined on a time-varying, finite grid. Other global techniques involved the use of Hermite expansions and Gauss-Hermite integration [36]. The first use of dynamic programming for nonlinear state estimation, an intrinsic element of hidden Markov modeling approaches, occurred in 1966 with Larson's modal trajectory approach [51]. The use of approximation techniques for nonlinear state estimation with continuous-time systems has been a more recent development. Kushner [47, 48] pioneered notable work on the continuous-time problem with his use of interpolated, finite-state Markov chains for approximating the continuous-time Wiener-Hopf equation, a nonlinear, partial differential equation, to propagate the *a posteriori* PDF through time.

3.6 State Estimation and Chaos

Several state estimation algorithms for chaotic systems have emerged in the past few years, primarily from the physics community. All of the algorithms are heuristic, with most exploiting known topological or ergodic properties of chaotic systems. One limitation of most of these algorithms, particularly those which exploit topological properties, is that they are effective only with large input signal-to-noise ratios (SNRs), typically those well in excess of 20 dB. In this section, we highlight those techniques of particular relevance to this thesis because of either their probabilistic or signal processing foundations.

In contrast to most other research on nonlinear state estimation, the research involving chaotic systems has focused on deterministic state equations, i.e., the absence of driving noise, fixed-interval smoothing, and scenarios involving unknown system dynamics. In particular, three, general problem scenarios have been investigated:

1. State estimation when the system dynamics are known, i.e., the chaotic map $\boldsymbol{f}$ in (3.5) or chaotic flow $\boldsymbol{F}$ in (3.7) is known.

2. State estimation when the system dynamics are unknown but a *clean* reference orbit is available. That is, in addition to an observation set $Y = \{\boldsymbol{y}(i)\}$, one is given an orbit segment, to which no noise has been added, which has been generated by the same chaotic system but with a different initial condition.

3. State estimation when the system dynamics are unknown and only the observation set is available for performing state estimation.

In addition, much of the state estimation work involving the second and third scenarios above has dealt with the use of embedding techniques [21] for reconstructing the system dynamics. With the most popular of these techniques, one assumes that only a single component $x_i(t)$ of the $\mathcal{N}$-dimensional state vector $\boldsymbol{x}(t)$ is available. With the scalar observations $x_i(t)$, one creates vectors $\boldsymbol{z}(t)$ which consist of time delayed samples of $\{x_i(t)\}$. Specifically, one chooses constants $T_1, T_2, \cdots, T_{\mathcal{M}-1}$, and defines the vectors $\boldsymbol{z}(t) \equiv [x_i(t), x_i(t-T_1), \cdots, x_i(t-T_{\mathcal{M}-1})]^T$. Usually, the T_i are chosen to be equally spaced, so that $T_i = i\tau$ for some constant τ. If τ and $\mathcal{M}$ are properly chosen, the attractor of the new system implicitly determined by $\boldsymbol{z}(t)$ will be an embedding of the original attractor. Equivalently, if $\boldsymbol{g}$ denotes the system implicitly defined by $\boldsymbol{z}(t)$, then there will be a diffeomorphism $\boldsymbol{h}$ such that $\boldsymbol{g} = \boldsymbol{h} \circ \boldsymbol{f}$.

Although embedding techniques have received much attention in recent years, they are still poorly understood. More importantly, common rules of thumb for choosing delays and dimensions of the new state vectors may not hold in the context of state estimation. The problem of state estimation with chaotic systems involving the original state vectors is a challenging problem unto itself, and some of the fundamental, poorly understood aspects of this problem may be obscured if embedding considerations are incorporated into the problem. In light of this, although the state estimation algorithms introduced in the next chapter are applicable to systems derived by embedding, we do not deal with embedded systems in this thesis.

One of the earliest reported state estimation algorithms for chaos is an iterative, approximate ML approach [22]. As noted in Section 3.3, if $Y = Y(M,N) \equiv \{\boldsymbol{y}(i)\}_{i=M}^{N}$, each noise term $\boldsymbol{v}(n)$ has zero mean and covariance matrix $\boldsymbol{R}$, and $\boldsymbol{f}$ is invertible, the log-likelihood

function, $\log p(Y(M,N);\boldsymbol{x}(n_0))$ for the restricted DTS/DTO scenario is given by

$$\begin{aligned}\log p(Y(M,N);\boldsymbol{x}(n_0)) &= C(M,N) \\ &-\frac{1}{2}\sum_{i=M}^{N}\left[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right]^T \boldsymbol{R}^{-1}\left[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right], \end{aligned} \tag{3.43}$$

where $C(M,N)$ is a normalizing constant. When $\boldsymbol{h}$ is the identity operator and $\boldsymbol{R}=\sigma^2\,\boldsymbol{I}_{\mathcal{N}}$, where σ is a real-valued constant and $\boldsymbol{I}_{\mathcal{N}}$ is the $(\mathcal{N}\times\mathcal{N})$-identity matrix, this reduces to

$$\log p(Y(M,N);\boldsymbol{x}(n_0)) = C(M,N) - \frac{1}{2\sigma^2}\sum_{i=M}^{N}\left\|\boldsymbol{y}(i)-\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\right\|^2. \tag{3.44}$$

In [22], each difference $\boldsymbol{y}(i)-\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))$ in the above expression is approximated by the first term in a Taylor series expansion

$$\begin{aligned}\boldsymbol{y}(i)-\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)) &= \boldsymbol{f}^{i-n_0}(\boldsymbol{f}^{-(i-n_0)}(\boldsymbol{y}(i)))-\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)) &(3.45)\\ &\approx \boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}(\boldsymbol{f}^{-(i-n_0)}(\boldsymbol{y}(i))-\boldsymbol{x}(n_0)) &(3.46)\end{aligned}$$

where $\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}$ is the derviative of $\boldsymbol{f}^{i-n_0}$ evaluated at $\boldsymbol{x}(n_0)$. With these approximations $\log p(Y(M,N);\boldsymbol{x}(n_0))$ reduces to

$$\begin{aligned}\log p(Y(M,N);\boldsymbol{x}(n_0)) &= C(M,N) \\ &-\frac{1}{2\sigma^2}\sum_{i=M}^{N}\|\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}(\boldsymbol{f}^{-(i-n_0)}(\boldsymbol{y}(i))-\boldsymbol{x}(n_0))\|^2 \end{aligned} \tag{3.47}$$

which is linear in the unknown state $\boldsymbol{x}(n_0)$ if $\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}$ is independent of $\boldsymbol{x}(n_0)$, which in general it is not. By assuming $\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}$ is in fact independent of $\boldsymbol{x}(n_0)$, one can obtain a closed-form expression for the ML estimate of $\boldsymbol{x}(n_0)$ by differentiating (3.47) and solving for $\boldsymbol{x}(n_0)$. Doing so yields

$$\begin{aligned}\boldsymbol{x}(n_0) &= \left[\sum_{i=M}^{N}\boldsymbol{D}^T\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\right]^{-1} \\ &\quad\times\sum_{i=M}^{N}\left[\boldsymbol{D}^T\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\boldsymbol{f}^{-(i-n_0)}(\boldsymbol{y}(i))\right] \end{aligned} \tag{3.48}$$

where $\boldsymbol{D}^T\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}$ denotes the transpose of $\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}$. The iterative ap-

proach used in [22] consists of using the current estimate of $\boldsymbol{x}(n_0)$ to calculate the derivative matrices $\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}$ and then evaluating the above expression to obtain the next estimate of $\boldsymbol{x}(n_0)$. As suggested in [22], this approximate ML state estimation approach is only effective if the approximated differences are small, or equivalently if the SNR is large. In Chapter 7, we introduce an estimator for a special class of piecewise linear maps for which an expression similar to (3.48) provides the exact ML state estimate. For this estimator, the appropriate derivatives $\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}$ are separately determined with a technique that uses hidden Markov modeling and Viterbi decoding.

Because of the form of the log-likelihood function given by (3.43), one can treat this function as an *energy function* and pose the problem of ML state estimation as a constrained optimization problem. Note that this form arises because of the additive, white, Gaussian assumption on the observation noise and the assumption of zero driving noise. In particular, (3.43) can be expressed

$$\begin{aligned}\log p(Y(M,N);\boldsymbol{x}(n_0)) &= C(M,N) \\ &-\frac{1}{2}\sum_{i=M}^{N}[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{z}(i))]^T\boldsymbol{R}^{-1}[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{z}(i))]+\sum_{i=M+1}^{N}\alpha_i[\boldsymbol{z}(i)-\boldsymbol{f}(\boldsymbol{z}(i-1))] \quad (3.49)\end{aligned}$$

where the $\{\boldsymbol{z}(i)\}$ are the unknowns and the $\{\alpha_i\}$ are Lagrange multipliers. For many chaotic maps, the chaotic nature of the dynamics precludes straightforward minimization of this cost function using standard optimization techniques.

A cost function similar in appearance to the one above, but fundamentally different and easier to minimize, is considered by Kostelich and Yorke in [45]. In the Kostelich and Yorke cost function, the $\{\alpha_i\}$ are treated as known, weighting constants rather than Lagrange multipliers. Treating the α_i in this way transforms the problem from a constrained minimization problem to a regularization problem with some similarity to those that arise in many image understanding applications [8]. In [45], an iterative, least squares approach is used to minimize the resulting cost function and simultaneously estimate the system dynamics.

An alternative approach to fixed-interval smoothing with chaotic maps is considered in [23, 35]. The technique discussed in [35] is applicable to two-dimensional, invertible chaotic maps with known system dynamics. The technique in [23] is similar but does not require knowledge of the system dynamics and instead uses a locally linear estimate of the

dynamics. The purported motivation of both approaches is the *Shadowing Lemma* [10, 11], a lemma which applies to a certain class of well-understood, nondissipative, chaotic systems known as Anosov diffeomorphisms.

The state estimation algorithms presented in [23, 35] are iterative algorithms in which one starts with a set of noise-corrupted observations $Y(0, N)$ for the restricted DTS/DTO scenario, with the observation function $\boldsymbol{h}$ in (3.6) given by the identity operator. For each iteration, one first estimates the stable and unstable manifolds [21] of the undriven system associated with the state $\boldsymbol{x}(n_0)$ at each time n_0 using the current state estimates. Next, two stable autoregressive processes, one running forward in time along the estimated stable manifold, the other running backward in time along the estimated unstable manifold, are used to propagate correction terms, with a pair of correction terms associated with the state at each time n_0. The pair of correction terms associated with the state at a given time are then added to the current estimate of the state yielding a new estimate. The overall objective of updating the states estimates is not to reduce the estimation error but to ensure that the new estimates more closely obey the system dynamics than the previous estimates. That is, if $\boldsymbol{f}$ denotes the system, $\hat{\boldsymbol{x}}_{new}(n_0)$ the new state estimate at time n_0, and $\hat{\boldsymbol{x}}_{old}(n_0)$ the previous estimate, then the correction terms for time n_0 are chosen with the goal that $\|\hat{\boldsymbol{x}}_{new}(n_0) - \boldsymbol{f}(\hat{\boldsymbol{x}}_{new}(n_0 - 1))\|$ be less than $\|\hat{\boldsymbol{x}}_{old}(n_0) - \boldsymbol{f}(\hat{\boldsymbol{x}}_{old}(n_0 - 1))\|$. Although not discussed in [23, 35], an inherent assumption in the algorithms is that the initial noise term $\boldsymbol{v}(0)$ has no component in the direction of the stable manifold of $\boldsymbol{x}(0)$ and the final noise term $\boldsymbol{v}(N)$ has no component in the direction of the unstable manifold of $\boldsymbol{x}(N)$.

Because the local stable and unstable manifolds of a chaotic orbit and the expansion rates associated with these manifolds are only applicable to infinitesimally small perturbations, the algorithms are potentially useful only when the initial SNR is large. In fact, experimental results suggest that the algorithms perform poorly below 20 dB, and even with large SNRs the algorithms often behave erratically.

As is shown in Chapters 4 and 5, the existence of both positive and negative Lyapunov exponents for dissipative, chaotic diffeomorphisms has a profound impact on achievable state estimator performance with these systems. In addition, the existence of these exponents and the associated stable and unstable manifolds is necessarily a fundamental consideration when developing practical, effective, probabilistic state estimators for dissipative chaotic systems.

Hidden Markov modeling approaches to state estimation with chaotic maps are presented in [56, 64]. With both approaches, the dynamics of the underlying, deterministic, chaotic system are approximated with a finite state hidden Markov model [74, 75] with the output of each state a constant vector in [56] and a Gaussian random vector in [64]. Neither approach requires knowledge of the actual system dynamics; but, both require availability of a noise-free reference orbit for estimating transition probabilities and output distributions.

The approach discussed in [56] is an iterative, fixed-interval, smoothing approach and applicable only with bounded, uniformly distributed observation noise. At each iteration, a local partitioning of state space is first performed around the current estimate of each orbit point. Each partition element corresponds to a state of the assumed underlying Markov chain. Next, a noise-free reference orbit is used to estimate transition probabilities between the states associated with one orbit point and those associated with the next. Finally, the Viterbi algorithm is used to determine the most likely state sequence (actually the centroids of the states) as determined by the noisy observations, transition probabilities, and assumption on the observation noise.

The approach in [64] uses a single set of states for all noisy orbit points. These states do not correspond to a partioning of the phase space. Instead, they correspond to a modeling of the dynamics of the chaotic system as a finite-state, first-order hidden Markov model, in which the noise-free output corresponding to each state is a Gaussian random vector. The state transition probabilities as well as the mean vectors and covariance matrices of the states are estimated from the clean reference orbit using the Baum-Welch re-estimation formula. Each observed, noisy, orbit point is modeled as the sum of the output of a state of the underlying hidden Markov model and an independent, random vector which represents the contribution of the observation noise. Heuristic ML and MMSE state estimators based on this model are introduced in [64]. The estimators perform reasonably well with small to moderate input SNRs, with performance for larger input SNRs strongly dependent on the number of states. The MMSE state estimator would be labeled a *global* technique with the taxonomy of nonlinear, suboptimal, MMSE state estimators discussed earlier in the chapter.

In Chapter 6, optimal and suboptimal detection and estimation algorithms based on hidden Markov modeling are introduced for a special class of one-dimensional, chaotic maps. The appropriateness and value of using HMMs in conjunction with these maps arises from

the fact that these maps give rise to homogeneous, finite-state Markov chains.

Chapter 4

State Estimation with Dissipative, Chaotic Maps

4.1 Introduction

In this chapter, we consider probabilistic state estimation with chaotic maps and introduce practical, suboptimal Maximum-Likelihood (ML) and Minimum-Mean-Squared-Error (MMSE) state estimators. We begin by discussing the experimentally observed properties of the likelihood function for the restricted DTS/DTO scenario given by (3.5) and (3.6) for the special case in which $\boldsymbol{h}$ is the identity operator and $\boldsymbol{f}$ is a two-dimensional, dissipative, chaotic map. We offer a nonrigorous explanation of these properties which is supported by the performance bound analysis in Chapter 5. We also introduce a simple, grid-based, ML estimator, elements of which are incorporated in the MMSE estimators introduced later in the chapter.

The chapter then focuses on MMSE state estimation and introduces two heuristic MMSE state estimators—one, a local estimator based on extended Kalman filtering, and the other, a global estimator that approximates the conditional mean integral given by (3.19) with a recursively calculable, finite summation. Performance results for both estimators are presented and compared. The results suggest that the local approach is more effective with larger input SNRs and the global approach more effective with smaller input SNRs.

In this chapter, our focus is on problem scenarios in which the input signal-to-noise ratio (SNR) is smaller than 20 dB, a heretofore relatively unexplored problem in the context of

chaos, but one that arises in real-world applications such as secure communication, potentially involving chaotic systems. All three problem scenarios listed in Section 3.6—known system dynamics, unknown system dynamics and availability of a noise-free reference orbit, unknown system dynamics and no availability of a noise-free reference orbit—are dealt with in this chapter, with emphasis placed on the second scenario. Whereas there are presently few practical applications involving chaos, it is unclear which of these scenarios has the most practical relevance.

4.2 ML State Estimation

In this section, we consider ML state estimation with dissipative, chaotic diffeomorphisms. Although two specific maps—the Henon and Ikeda maps—are emphasized, many of the results apply to other dissipative, chaotic maps as well, even those with state vector dimensions greater than two. The section first qualitatively analyzes the properties of the likelihood function that arises with the DTS/DTO scenario, in part to motivate the approximate ML state estimator introduced later in the section. The specific problem considered is that of estimating $\boldsymbol{x}(n_0)$, the state of the dissipative, chaotic map $\boldsymbol{f}$ at time n_0, given the observation set $Y \equiv Y(M,N) = \{\boldsymbol{y}(i)\}_{i=M}^{N}$ for the restricted DTS/DTO scenario with state and observation equations (3.5) and (3.6), respectively.

4.2.1 Properties of the Likelihood Function

As shown in Chapter 3, the log-likelihood function, $\log p(Y;\boldsymbol{x}(n_0))$, has the following form for the restricted DTS/DTO scenario:

$$\begin{aligned}\log p(Y;\boldsymbol{x}(n_0)) &= C(M,N) \\ &\quad -\frac{1}{2}\sum_{i=M}^{N}\left[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right]^T \boldsymbol{R}^{-1}\left[\boldsymbol{y}(i)-\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right], \end{aligned} \tag{4.1}$$

where $C(M,N)$ is a normalizing constant and $\boldsymbol{R}$ is the covariance matrix of the observation noise. Note that (4.1) is a sum of weighted, squared-error terms in which each error is the difference between an observation vector $\boldsymbol{y}(i)$ and the corresponding transformed state vector $\boldsymbol{x}(n_0)$, and the weighting matrix is the inverse of the covariance matrix of the observation noise. Whereas the likelihood function has a fundamental role in probabilistic

state estimation, an understanding of the basic properties of this function for a given state estimation problem is useful when developing practical state estimators for that problem. With this in mind, we now provide a nonrigorous analysis supported by experimental results of the properties of $p(Y;\boldsymbol{x}(n_0))$ for the special case in which $\boldsymbol{h}$ is the identity operator and $\boldsymbol{R}$ is a diagonal, positive definite matrix with diagonal elements $\{\sigma_i^2\}_{i=1}^{\mathcal{N}}$, where $\mathcal{N}$ is the dimension of $\boldsymbol{x}(n_0)$. Although the simulations involve only two-dimensional maps, the properties apply to higher-dimensional maps as well. Both the nonrigorous analysis presented here and the performance bound analysis presented in Chapter 5 suggest a fundamental influence of the system Lyapunov exponents on the likelihood function for chaotic systems.

We first consider the case in which $M = n_0 - m$ and $N = n_0$ in (4.1) for some positive integer m, which corresponds to all observations occurring before or at the time of interest n_0. Figures 4-1 (a) and (b) are contour plots of the likelihood function $p(Y(n_0-10,n_0);\boldsymbol{x}(n_0))$ for a fixed time n_0 for the Henon and Ikeda maps with an input SNR of 6 dB. The figures

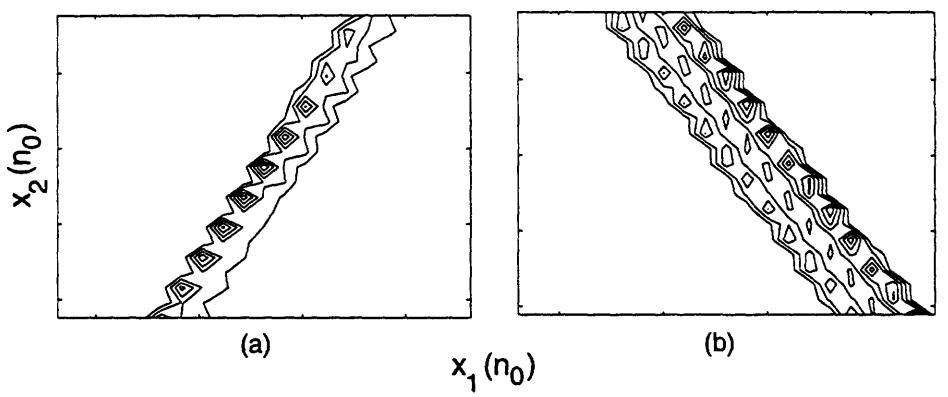

Figure 4-1: Contour plot of $p(Y(n_0-10,n_0);\boldsymbol{x}(n_0))$ as a function of $\boldsymbol{x}(n_0) = [x_1(n_0),x_2(n_0)]^T$. (a) Henon map; (b) Ikeda map.

depict the relative values of the likelihood function $p(Y(n_0-10,n_0);\boldsymbol{x}(n_0))$ as a function of $\boldsymbol{x}(n_0)$ for a fixed set of observations. In other words, a set of observations $Y(n_0-10,n_0)$ was generated using the orbit segment $\{\boldsymbol{x}(i)\}_{i=n_0-10}^{n_0}$ with $\boldsymbol{x}(n_0) = \boldsymbol{x}_{n_0}$. The likelihood function was then evaluated for various values of $\boldsymbol{x}(n_0)$ using this set of observations, and the relative likelihood values shown in the figures. The center point in each figure is the relative likelihood value corresponding to $\boldsymbol{x}_{n_0}$ and the other values are those for a rectangular grid of points (i.e., values of $\boldsymbol{x}(n_0)$) centered at $\boldsymbol{x}_{n_0}$ with a grid spacing of .002. In the figures, the nesting of contours indicates increasing values of the likelihood function. Alternative graphical representations of the data are provided in Figures 4-2 (a) and (b) which are mesh

plots of the same data used for Figures 4-1 (a) and (b), respectively. Whereas ML state

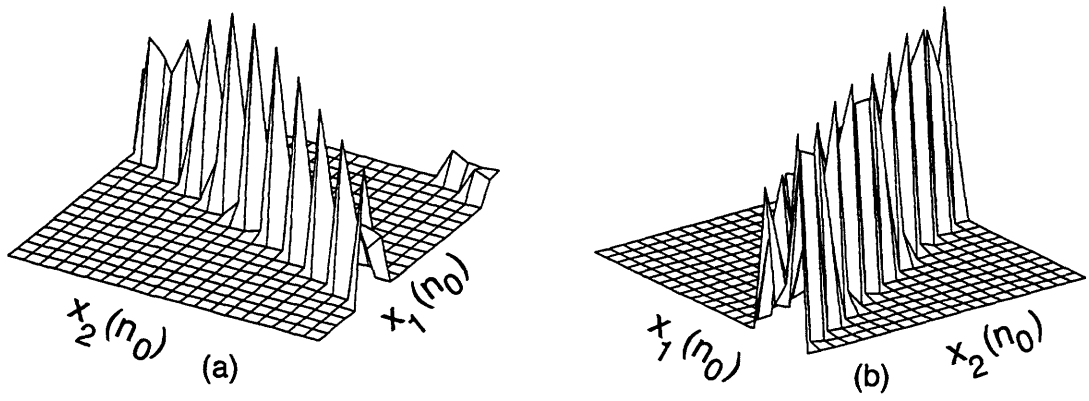

Figure 4-2: Mesh plot of $p(Y(n_0-10,n_0);\boldsymbol{x}(n_0))$ as a function of $\boldsymbol{x}(n_0)+[x_1(n_0),x_2(n_0)]^T$. (a) Henon map; (b) Ikeda map.

estimation would entail choosing the value of $\boldsymbol{x}(n_0)$ for which $p(Y(n_0-10,n_0);\boldsymbol{x}(n_0))$ is largest, the ridge-like property of the likelihood function in the figures suggest that the ML estimate of $\boldsymbol{x}_{n_0}$, the actual value of $\boldsymbol{x}(n_0)$, may be a very poor estimate for both maps. The results are similar with observations for times at or after n_0. Figures 4-3 (a) and (b) are contour plots of $p(Y(n_0,n_0+15);\boldsymbol{x}(n_0))$ as a function of $\boldsymbol{x}(n_0)$ for the Henon and Ikeda maps.

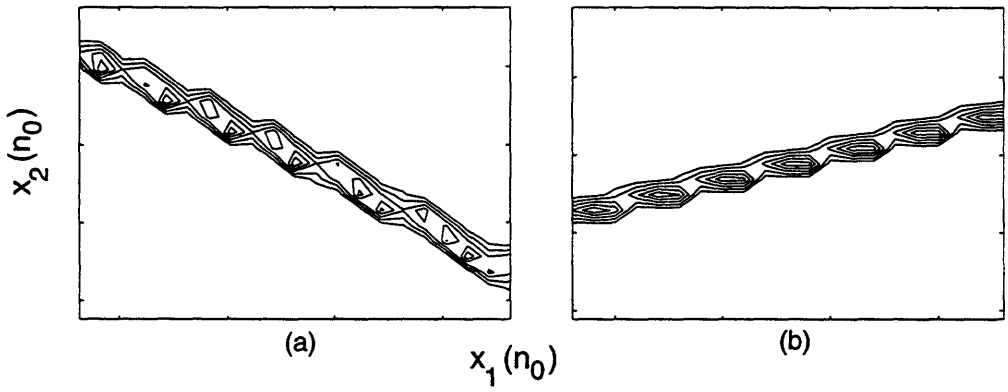

Figure 4-3: Contour plot of $p(Y(n_0,n_0+15);\boldsymbol{x}(n_0))$ as a function of $\boldsymbol{x}(n_0)$: (a) Henon map; (b) Ikeda map.

A nonrigorous, first-order analysis helps explain this interesting property of the likelihood function. Consider a single term of the sum in (4.1):

$$S_i(\boldsymbol{x}(n)) = \left[\boldsymbol{y}(i) - \boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\right]^T \boldsymbol{R}^{-1} \left[\boldsymbol{y}(i) - \boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\right]. \tag{4.2}$$

For a small deviation $\boldsymbol{\delta}$ from the actual value of $\boldsymbol{x}(n_0)$, the following relation holds to first-order:

$$\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)+\boldsymbol{\delta}) \approx \boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)) + \boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}, \tag{4.3}$$

where $\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}$ denotes the derivative of $\boldsymbol{f}^{i-n_0}$ with respect to $\boldsymbol{x}(n_0)$. Substituting $\boldsymbol{x}(n_0)+\boldsymbol{\delta}$ for $\boldsymbol{x}(n_0)$ in (4.2) and also substituting (4.3) in (4.2) yields

$$\begin{aligned} S_i(\boldsymbol{x}(n_0)+\boldsymbol{\delta}) \approx & \left[\boldsymbol{y}(i) - \boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)) - \boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}\right]^T \\ & \times \boldsymbol{R}^{-1}\left[\boldsymbol{y}(i) - \boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)) - \boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}\right], \end{aligned} \tag{4.4}$$

which because of the diagonal assumption on $\boldsymbol{R}$ reduces to

$$S_i(\boldsymbol{x}(n_0)+\boldsymbol{\delta}) \approx \sum_{k=1}^{\mathcal{N}} \frac{\left\{v_k(i) - \left[\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}\right]_k\right\}^2}{\sigma_k^2} \tag{4.5}$$

where $v_k(i)$ denotes the k^{th} element of the noise vector $\boldsymbol{v}(i)$ and $\left[\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}\right]_k$ denotes the k^{th} element of the vector $\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}$.

Since $\boldsymbol{v}(i)$ is zero-mean, the expected value of $S_i(\boldsymbol{x}(n_0)+\boldsymbol{\delta})$ is given by

$$E\{S_i(\boldsymbol{x}(n_0)+\boldsymbol{\delta})\} \approx \sum_{k=1}^{\mathcal{N}} \frac{\sigma_k^2 + \left[\boldsymbol{D}\{\boldsymbol{f}^{(i-n_0)}(\boldsymbol{x}(n_0))\}\boldsymbol{\delta}\right]_k^2}{\sigma_k^2}, \tag{4.6}$$

where E is the expectation operator. The log-likelihood function and consequently the likelihood function is large if each term in the sum is small. As such, the value of the likelihood function for the perturbed state $\boldsymbol{x}(n_0)+\boldsymbol{\delta}$ depends on the magnitudes of the vectors

$$\boldsymbol{D}\{\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}, \quad i = M, \cdots, N. \tag{4.7}$$

As discussed in Chapter 2, an $\mathcal{N}$-dimensional chaotic map has $\mathcal{N}$ Lyapunov exponents, and associated with each exponent and point on the attractor is a linear subspace of $R^{\mathcal{N}}$ for which the logarithm of the long-term, averaged growth rate of infinitesimal perturbations along most of that subspace is given by the Lyapunov exponent. More precisely, if $\lambda_1 > \lambda_2 > \cdots > \lambda_{\mathcal{N}}$ denote the ordered Lyapunov exponents not repeated by multiplicity and $E^j_{\boldsymbol{x}(n_0)}$ denotes the subspace of $R^{\mathcal{N}}$ associated with $\boldsymbol{x}(n_0)$ and all Lyapunov exponents less

than or equal to λ_j, then the following holds [21]:

$$\lim_{n\to\infty} \frac{1}{n} \log \|\boldsymbol{D}\{\boldsymbol{f}^n(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}\| = \lambda_j \tag{4.8}$$

for any unit vector $\boldsymbol{\delta} \in E^j_{\boldsymbol{x}(n_0)} \setminus E^{j+1}_{\boldsymbol{x}(n_0)}$, where $\|\cdot\|$ denotes the Euclidean norm and $E^j_{\boldsymbol{x}(n_0)} \setminus E^{j+1}_{\boldsymbol{x}(n_0)}$ denotes the set formed by taking $E^j_{\boldsymbol{x}(n_0)}$ and removing the subspace $E^{j+1}_{\boldsymbol{x}(n_0)}$.

As mentioned in Chapter 2, a frequently used but often poor approximation is that if $\boldsymbol{\delta}$ is a small perturbation, not necessarily infinitesimally small, and $\boldsymbol{\delta} \in E^j_{\boldsymbol{x}(n_0)} \setminus E^{j+1}_{\boldsymbol{x}(n_0)}$ then $\|\boldsymbol{D}\{\boldsymbol{f}^k(\boldsymbol{x}(n_0))\}\,\boldsymbol{\delta}\| \approx \|\boldsymbol{\delta}\| \exp(k\lambda_j)$. Although this approximation is often poor, its implications often are nonetheless valid because of the close relation between each linear subspace $E^j_{\boldsymbol{x}(n_0)}$ and a nonlinear counterpart known as a differentiable manifold. In particular, tangent to each of the linear subspaces $E^j_{\boldsymbol{x}(n)}$ corresponding to a negative Lyapunov exponent is a nonlinear differentiable manifold, with the orbits of points along the manifold exhibiting the expected, scaling behavior. That is, if $W^j_{\boldsymbol{x}(n_0)}$ is the nonlinear manifold associated with the linear manifold $E^j_{\boldsymbol{x}(n_0)}$, then for any point $\boldsymbol{x}(n_0) + \boldsymbol{\delta} \in W^j_{\boldsymbol{x}(n)}$, the following is true [21]:

$$\|\boldsymbol{f}^k(\boldsymbol{x}(n)) - \boldsymbol{f}^k(\boldsymbol{x}(n_0) + \boldsymbol{\delta})\| \leq C \exp(k\,\lambda_j). \tag{4.9}$$

(These nonlinear manifolds are defined only for negative Lyapunov exponents).

One implication of this scaling behavior along these nonlinear manifolds is that the magnitudes of the vectors given by (4.7) are smallest for perturbations $\boldsymbol{\delta}$ along the nonlinear manifold associated with $\boldsymbol{x}_{n_0}$ corresponding to the smallest Lyapunov exponent of $\boldsymbol{f}$ if $i - n_0 > 0$ and the smallest Lyapunov exponent of $\boldsymbol{f}^{-1}$ if $i - n_0 < 0$. Note that the Lyapunov exponents of $\boldsymbol{f}^{-1}$ are the negatives of the Lyapunov exponents of $\boldsymbol{f}$.

This implication suggests that the ridge-like regions of large likelihood values in Figures 4-1 (a) and (b) correspond to the nonlinear manifold associated with the smallest Lyapunov exponent of $\boldsymbol{f}^{-1}$, and the ridge-like regions of large likelihood values in Figures 4-3 (a) and (b) correspond to the nonlinear manifold associated with the smallest Lyapunov exponent of $\boldsymbol{f}$. The reasoning behind this conclusion is the following. The results depicted in Figures 4-1 (a) and (b) are those for the case in which all observations occur before or at the time of interest so that $i - n_0 \leq 0$ in (4.6). Thus, the magnitudes of the vectors given by (4.7) are smallest for perturbed state vectors $\boldsymbol{x}_{n_0} + \boldsymbol{\delta}$ along the nonlinear manifold associated with $\boldsymbol{x}_{n_0}$ corresponding to the smallest Lyapunov exponent of $\boldsymbol{f}^{-1}$. Similarly, the results depicted in

Figures 4-3 (a) and (b) are those for the case in which all observations occur at or after the time of interest so that $i - n_0 \geq 0$ in (4.6). Thus, the magnitudes of the vectors given by (4.7) are smallest for perturbed state vectors along the nonlinear manifold associated with $\boldsymbol{x}_{n_0}$ corresponding to the smallest Lyapunov exponent of $\boldsymbol{f}$. Since the likelihood function is largest at points $\boldsymbol{x}(n_0)$ for which the vectors in (4.7) have the smallest magnitudes, the implication is that the ridge-like regions in the figures correspond to the nonlinear manifolds associated with $\boldsymbol{x}_{n_0}$ corresponding to these smallest Lyapunov exponents.

If correct, this conclusion applies to higher-dimensional chaotic systems as well. However, in contrast to two-dimensional, dissipative, chaotic diffeomorphisms which always have one positive and one negative Lyapunov exponent, higher-dimensional, dissipative, chaotic diffeomorphisms may have several negative and/or several positive Lyapunov exponents. As a result, for higher-dimensional systems if other Lyapunov exponents are comparable in size to the smallest exponent, then the likelihood function corresponding to observations at times before or at n_0 might be large for values of $\boldsymbol{x}(n_0)$ for a higher-dimensional manifold associated with $\boldsymbol{x}_{n_0}$. A similar result applies if other Lyapunov exponents are comparable in size to the largest Lyapunov exponent and all observations are at times at or after n_0.

The question arises as to the behavior of the likelihood function which includes observations at times both before and after n_0. Figures 4-4 (a) and (b) are contour plots of this likelihood function $p(Y(M,N);\boldsymbol{x}(n_0))$ for the Henon map with two sets of values for M and N, satisfying $M < n < N$ and Figures 4-5 (a) and (b) are analogous contour plots for the Ikeda map. As with the earlier contour plots, the figures depict the likelihood function

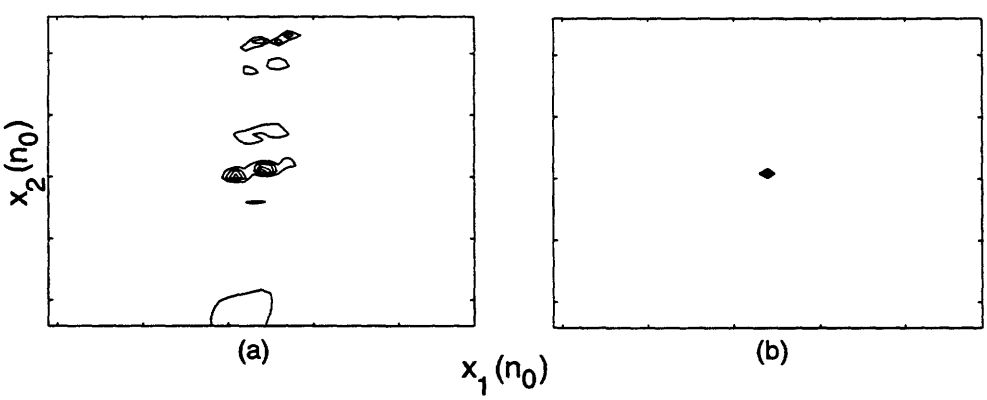

Figure 4-4: Contour plot of $p(Y(M,N);\boldsymbol{x}(n_0))$ as a function of $\boldsymbol{x}(n_0)$ for the Henon map. (a) $M = n_0 - 3$ and $N = n_0 + 12$; (b) $M = n_0 - 5$ and $N = n_0 + 20$.

evaluated at a a rectangular grid of values for $\boldsymbol{x}(n_0)$ centered at $\boldsymbol{x}_{n_0}$ with a fixed observation

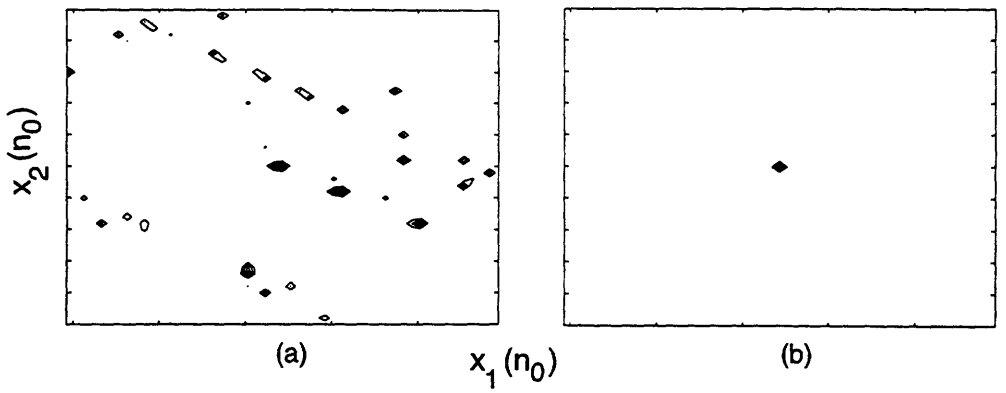

Figure 4-5: Contour plot of $p(Y(M,N);\boldsymbol{x}(n_0))$ as a function of $\boldsymbol{x}(n_0)$ for the Ikeda map. (a) $M = n_0 - 3$ and $N = n_0 + 15$; (b) $M = n_0 - 6$ and $N = n_0 + 25$.

set Y generated with $\boldsymbol{x}_{n_0}$. As indicated by the figures, the likelihood function is multimodal with only a few observations, but rapidly becomes impulse-like as the number of past and future observations increases.

Figures 4-6 (a) and (b) depict analogous results for a slightly shifted rectangular grid of values for $\boldsymbol{x}(n_0)$. In particular, the grid used for the figures was centered at the perturbed state $\boldsymbol{x}(n)+[1.5\times10^{-4}, 1.5\times10^{-4}]^T$ and $\boldsymbol{x}_{n_0}$ was not a grid point. The grid spacing in both figures as in the earlier figures is .002. A comparison of Figures 4-4 (b) and 4-6 (a) and a

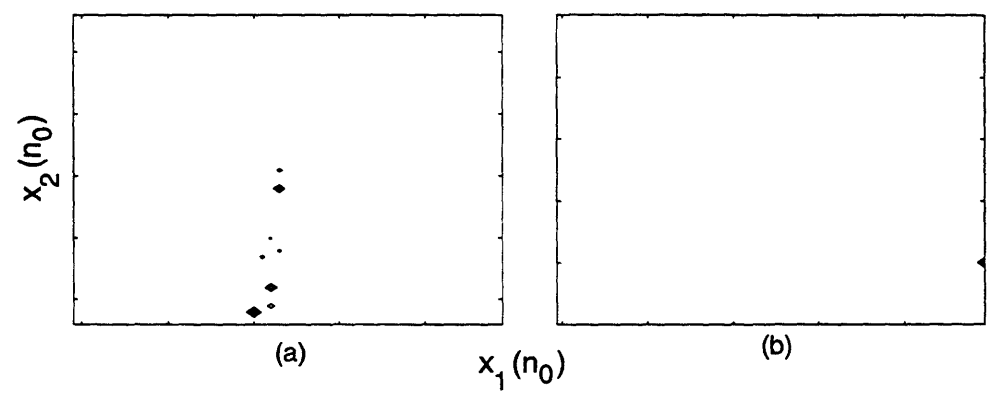

Figure 4-6: Contour plot of $p(Y(M,N);\boldsymbol{x}(n_0))$ as a function of $\boldsymbol{x}(n_0)$ with value for $\boldsymbol{x}_{n_0}$ not shown. (a) Henon map with $M = n_0 - 5$ and $N = n_0 + 20$; (b) Ikeda map with $M = n_0 - 6$ and $N = n_0 + 25$

comparison of Figures 4-5 (b) and 4-6 (b) reveals an extreme sensitivity of the likelihood function to the grid spacing when the numbers of past and future observations are not small. This sensitivity is shown to have an important influence on the performance of an approximate ML state estimator introduced in the next section.

The impulse-like nature of the likelihood function $p(Y(M,N);\boldsymbol{x}(n_0))$ for invertible, dissipative, chaotic systems with observations at both past and future times suggests that the ML estimate of $\boldsymbol{x}_{n_0}$ becomes increasingly accurate as the numbers of observations for both past and future times increases. However, several practical considerations thwart maximization of the likelihood function. First, the likelihood function for a dissipative, chaotic map is highly nonlinear and as a consequence there is generally no closed-form expression for the ML estimate of $\boldsymbol{x}_{n_0}$. Second, the likelihood function generally has multiple local minima for small numbers of observations; this property coupled with the the impulse-like nature of the likelihood function for larger numbers of observations precludes the straightforward use of standard, nonlinear, optimization techniques for numerically maximizing the likelihood function. Third, the inverse of an attractor is a repeller. That is, whereas a dissipative, chaotic system $\boldsymbol{f}$ shrinks volumes in the basin of attraction of an attractor, the inverse system $\boldsymbol{f}^{-1}$ expands volumes and is thus inherently unstable. Because of this instability, the orbits of $\boldsymbol{f}^{-1}$ generated by most initial conditions on the attractor are unbounded. As a result, it is generally difficult to obtain accurate estimates of the backward orbit of a point, i.e., the orbit with the given point as the final condition. Finally, the use of standard optimization techniques requires knowledge of the system dynamics which may not always be available.

Nonetheless, potentially effective, approximate, ML state estimation is possible with dissipative, chaotic diffeomorphisms even when the system dynamics are unknown. The next section introduces a simple state estimator which circumvents the above-mentioned problems by exploiting the topological transitivity property of chaotic systems.

4.2.2 An Approximate ML State Estimator

In this section, we introduce an approach for practical, suboptimal, ML state estimation with dissipative, chaotic maps. We consider a more general problem than that of estimating $\boldsymbol{x}(n_0)$, the state at time n_0. In particular, we consider the problem of estimating the $(N+1)$-point orbit segment $\{\boldsymbol{f}^i(\boldsymbol{x}(0))\}_{i=0}^{N}$ given the observation set $Y(0,N)$. In theory, the two problems are equivalent since the system dynamics are deterministic. That is, if $\hat{\boldsymbol{x}}_{ML}(n_0)$ denotes the ML estimate of $\boldsymbol{x}(n_0)$ for a given observation set $Y(0,N)$, then $\boldsymbol{f}^{n-n_0}(\hat{\boldsymbol{x}}_{ML}(n_0))$ denotes the ML estimate of $\boldsymbol{x}(n)$ for arbitrary time n for the same observation set. However, for the approximate ML approach considered in this section, experimental performance re-

sults are generally superior when the state is estimated separately at each time n. This appears to be a consequence of the simultaneous presence of both positive and negative Lyapunov exponents with a dissipative, chaotic map $\boldsymbol{f}$ which leads to amplification of estimation errors in state estimates when acted upon by $\boldsymbol{f}$ or $\boldsymbol{f}^{-1}$. In other words, if $\hat{\boldsymbol{x}}(n)$ denotes an estimate of $\boldsymbol{x}(n)$ with nonzero, squared estimation error ϵ^2, then the squared estimation errors for the estimates $f(\hat{\boldsymbol{x}}(n))$ and $f^{-1}(\hat{\boldsymbol{x}}(n))$ of $\boldsymbol{x}(n+1)$ and $\boldsymbol{x}(n-1)$, respectively, are generally larger than ϵ^2. This difference between theory and practice results both from nonoptimality of the estimator and computer round-off error. In Chapter 7, we introduce an optimal ML state estimator for a class of expanding, one-dimensional maps with a similar difference between theory and practice, but with the difference attributable in this case solely to computer round-off error.

An underlying requirement of the estimation approach introduced here is that the orbit segment $\{f^i(\boldsymbol{x}(0))\}_{i=0}^{N}$ to be estimated lies on a chaotic attractor. This requirement is assumed to hold in the following discussion. Also, unless stated otherwise, the distance between two orbit points refers to the Euclidean distance between the points and the distance between orbit segments refers to the sum of the Euclidean distances between corresponding points on the orbits divided by the number of segment points.

One numerical method for ML parameter estimation, the method we use here, is that of grid search. With this method, one evaluates the likelihood function at a finite set of parameter values and uses the parameter value for which likelihood function is largest as the parameter estimate. With respect to the ML state estimation problem of interest here, the parameter corresponds to $\boldsymbol{x}_{n_0}$, the actual unknown state at time n_0 where $n_0 \in [0, N]$. An important consideration with this estimation method is the selection of an appropriate set of possible parameter values, which for the state estimation problem correspond to trial values for $\boldsymbol{x}_{n_0}$. Complicating the task of choosing an appropriate set of trial values for the state estimation problem are two practical problems noted earlier which arise with dissipative, chaotic systems—the difficulty in generating accurate backward orbits for points and the impulse-like nature of the likelihood function with large numbers of observations at past and future times. The first problem is relevant because backward orbit segments for the trial state values are used in the likelihood function if $n_0 \neq 0$, as indicated by (4.1). The second problem is relevant if N is large. One simple, albeit suboptimal way to circumvent this second problem is to only use observations for times near n_0 in the likelihood function.

However, the first problem remains since observations for times both before and after n_0 are needed in the likelihood function for it to avoid the undesirable ridge-like behavior which occurs when all observations are for times at or after n_0. A desirable situation would be one in which an appropriate set of trial values was available for which the orbit, or at least an adequately sized orbit segment, containing each of the values was known as well.

Since a chaotic system is topologically transitive, most orbits on an attractor have points which come arbitrarily close to all other points on the attractor. In light of this and the underlying requirement that the unknown orbit segment lies on the attractor, the points on almost any sufficiently long orbit segment on the attractor are useful trial values for grid-based ML state estimation, since the orbit corresponding to these points is known and a subset of them pass arbitrarily close to $\boldsymbol{x}_{n_0}$.

Combining these related ideas leads to the following grid-based, approximate, ML state estimator for estimating the state at each time $n_0 \in [0, N]$:

Approximate ML State Estimator

1. Generate a reference orbit with arbitrary initial condition (with the possible exception of a set of measure zero) in the basin of attraction of the attractor. Let $\{\boldsymbol{o}_i\}$ denote the set of orbit points.

2. Evaluate the likelihood function $p(Y(n_0 - m, n_0 + r); \boldsymbol{o}_i)$ for the given observation set and each reference orbit point $\boldsymbol{o}_i$.

3. Choose as the ML state estimate at time n_0, the reference orbit point which maximizes the likelihood function.

The algorithm requires specification of the reference orbit length as well as the constants m and r, which determine the number of observations at past and future times to use in the likelihood function. Although discovering an optimal technique for selecting these values remains an elusive goal, there are several practical rules-of-thumb that are appropriate. In particular, for a given, desired performance level (e.g., mean-squared error), the orbit length must be such that at least one orbit point passes sufficiently close to the actual state $\boldsymbol{x}_{n_0}$ for each $n_0 \in [0, N]$ so that the desired performance level is achievable. Also, the constants m and r define a window of observation times and should be chosen so that orbit segments passing through two neighboring points remain close, on a point by point basis, over the

observation window. In other words, because the system dynamics are deterministic and invertible by assumption, the orbits passing through two neighboring points on the attractor remain close for at least a few time steps into the future and a few time steps into the past as well. A fundamental aspect of the algorithm is that its use does not actually require knowledge of the system dynamics, but only the availability of a reference orbit on the attractor.

As implicitly suggested by the expression for $\log p(Y(M,N);\boldsymbol{x}(n_0))$ given by (4.1), in the special case that $\boldsymbol{h}$ is the identity operator, the estimator is simply an orbit matcher which chooses as the state estimate at time n_0, the reference orbit point for which the corresponding orbit segment as determined by m and r most closely matches in a weighted least squares sense the set of observations $Y(n_0 - m, n_0 + r)$, with the weights determined by $\boldsymbol{R}^{-1}$. Equivalently, noise-free reference orbit segments are matched to noise-corrupted orbit segments. When $\boldsymbol{h}$ is not the identity operator, the estimator is a transformed orbit matcher in which noise-free orbit segments, with each point transformed by $\boldsymbol{h}$, are matched to noise-corrupted orbit segments, also with each point transformed by $\boldsymbol{h}$, prior to the noise corruption.

This approximate ML state estimator was tested on the Henon and Ikeda maps with the identity operator used for $\boldsymbol{h}$ in (3.6) and a diagonal matrix used for the noise covariance matrix $\boldsymbol{R}$. Figures 4-7 (a) and (b) depict the SNR gain as function of input SNR, with the curves parameterized by the number of points in the reference orbit segment. Each of the plotted points represents the average improvement in SNR for the two components of the state vector, obtained by estimating 2000 consecutive orbit points. As indicated

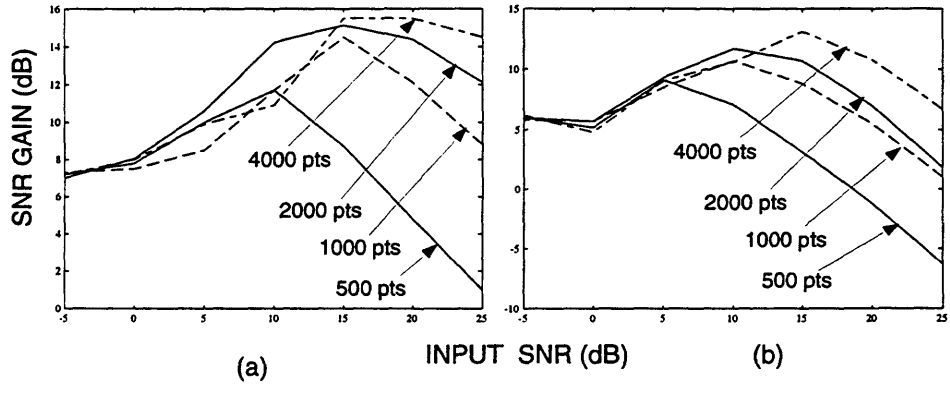

Figure 4-7: Performance results for approximate ML estimator with differently sized reference orbits and $(m,r) = (4,4)$. (a) Henon map; (b) Ikeda map.

by the figures, for input SNRs of 10 dB and above performance improves as the size of the reference orbit increases, whereas for input SNRs smaller than 10 dB the size of the reference orbit has little influence on performance (for reference orbit sizes in excess of 500 points). In addition, as is typical of ML-type estimators for nonlinear estimation problems, the estimator is most effective with larger input SNRs. The decrease in performance as the input SNR increases from 15 to 25 dB is attributable to the reference orbit size. Figures 4-8 (a) and (b) depict similar results, but with the curves now parameterized by the number of past and future observations (m, r) used in the likelihood function. The figures suggest

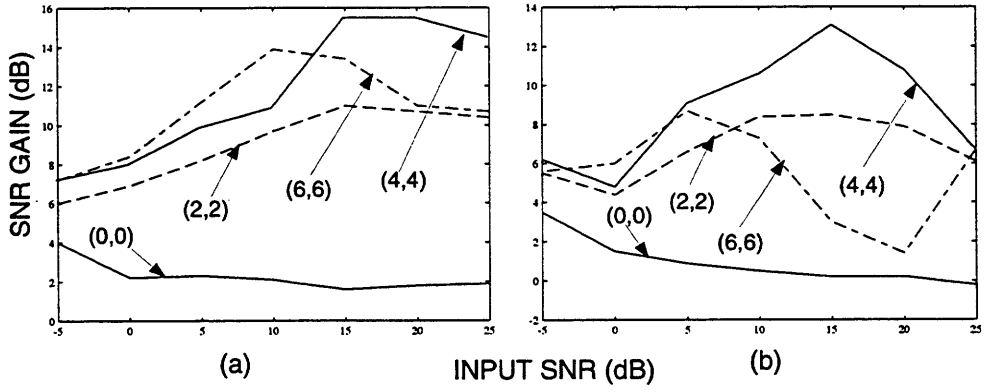

Figure 4-8: Performance results for approximate ML estimator with different numbers of observations (m, r) used in the likelihood function and a 4000-point reference orbit. (a) Henon map; (b) Ikeda map.

that performance improves initially as the values of m and r increase from zero, but then deteriorates as m and r exceed certain values. The figures also suggest that for both maps, no single parameter pair (m, r) achieves the best performance with all input SNRs.

4.2.3 Extensions

A straightforward extension to the approximate ML estimator yields performance results comparable to those of the original algorithm, but more importantly renders the algorithm potentially useful even when no reference orbit is available. When a reference orbit is available, the extension consists of using the average, possibly weighted, of the P reference orbit points for which the likelihood function is largest as the state estimate at a given time. (An alternative extension not pursued here would be to use the average of all reference orbit points for which the likelihood function exceeds a suitably chosen threshold as the state estimate). The original algorithm corresponds to the special case of the extension for

which P equals 1. As shown later in this section, when no reference orbit is available, the output transformation $\boldsymbol{h}$ is the identity operator, and the number of observations $N+1$ is sufficiently large, one obtains a nonnegligible improvement in SNR with this extension to the approximate ML estimator.

Figures 4-9 and 4-10 depict the performance results obtained with this averaging estimator for $P = 5$ and simple averaging used for the state estimate at each time $n_0 \in [0, N]$. The parameterization of the curves in Figures 4-9 and 4-10 is the same as in Figures 4-7 and 4-8, respectively. A comparison of the plotted results with those shown in Figures 4-7

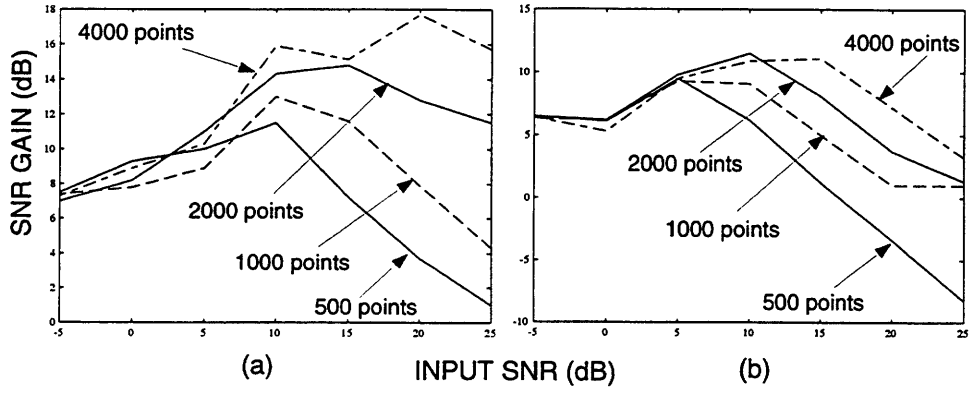

Figure 4-9: Performance results for averaging estimator with differently sized reference orbits, $(m, r) = (4, 4)$, and averaging over 5 reference orbit points. (a) Henon map; (b) Ikeda map.

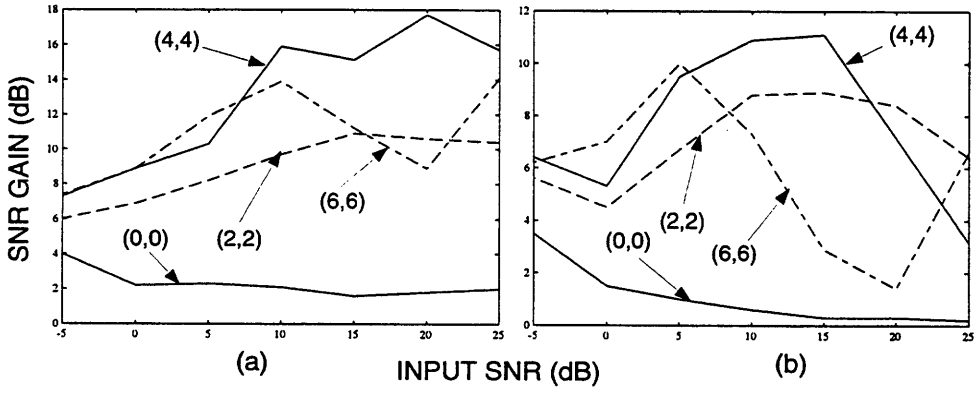

Figure 4-10: Performance results for averaging estimator with different numbers of observations (m, r) used in the likelihood function, a 4000-point reference orbit, and averaging over 5 reference orbit points. (a) Henon map; (b) Ikeda map.

and 4-8 indicate that averaging over 5 reference orbit points offers no additional SNR gain over the use of a single reference orbit point as the state estimate. Computer experiments involving averaging over other numbers of reference orbit points also yielded no additional

SNR gain.

However, averaging has a useful role in the so-called "self-cleaning" problem, the problem in which no reference orbit is available, the system dynamics are unknown, and the output transformation $\boldsymbol{h}$ is the identity operator, so that each observation $\boldsymbol{y}(n_0)$ is simply a noise-corrupted version of $\boldsymbol{x}(n_0)$, the state at that time. We assume with some loss of generality that the observation noise covariance matrix $\boldsymbol{R}$ is known as well. By using the observation set as a reference orbit, we can apply the averaging estimator to this problem. In other words, we let the observation set $Y(0,N)$ have two roles—one, the original role in the likelihood function $p(Y(n_0-m,n_0+r);\boldsymbol{o}_i)$ as the fixed observation subset $Y(n_0-m,n_0+r)$, and the other as the set of reference orbit points $\{\boldsymbol{o}_i\}$ over which the likelihood function is maximized, where $\log p(Y(n_0-m,n_0+r);\boldsymbol{o}_i)$ is now formally defined as

$$\log p(Y(n_0-m,n_0+r);\boldsymbol{o}_i) =$$

$$C(m,r)-\frac{1}{2}\sum_{j=-m}^{r}[\boldsymbol{y}(n_0+j)-\boldsymbol{o}_{i+j}]^T\boldsymbol{R}^{-1}[\boldsymbol{y}(n_0+j)-\boldsymbol{o}_{i+j}] \tag{4.10}$$

$$=C(m,r)-\frac{1}{2}\sum_{j=-m}^{r}[\boldsymbol{y}(n_0+j)-\boldsymbol{y}(i+j)]^T\boldsymbol{R}^{-1}[\boldsymbol{y}(n_0+j)-\boldsymbol{y}(i+j)] \tag{4.11}$$

where $C(m,r)$ is a normalizing constant. Strictly speaking, $p(Y(n_0-m,n_0+r);\boldsymbol{o}_i)$ is not a likelihood function for the underlying system model when $\boldsymbol{o}_i=\boldsymbol{y}(i)$ because $Y(0,N)$ is not an orbit segment. As such, referring to this estimation approach as ML state estimation is inappropriate.

It follows from (4.11) that $\log p(Y(n_0-m,n_0+r);\boldsymbol{y}(n_0))=0$ so that the likelihood function is maximized at time n_0 by $\boldsymbol{y}(n_0)$. Thus, choosing the observation which maximizes the likelihood function at each time n_0 as the state estimate would yield no improvement in SNR.

Figures 4-11 (a) and (b) depict the performance results obtained by applying this self-cleaning, averaging estimator to the Henon and Ikeda maps. Each plotted point is the average improvement in SNR for a 2000 point observation set and $(m,r)=(4,4)$. In contrast to the curves in Figures 4-9 and 4-10, the curves in Figures 4-11 (a) and (b) are parameterized by the number of neighboring observations, as determined by orbit matching, which were used for averaging in order to estimate the state at each time n_0. The actual observation at each time was not included in the average for that time. As suggested by the

figures, averaging over neighboring points has a beneficial role in this situation and there is a nonnegligible SNR improvement for each input SNR for both maps. The performance

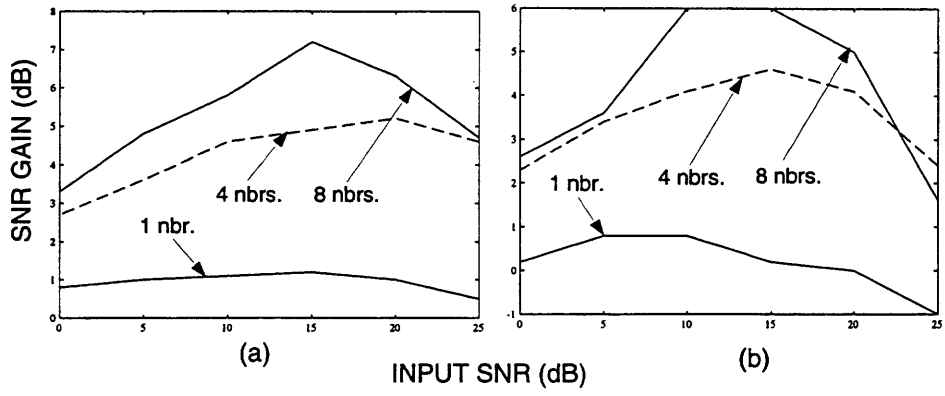

Figure 4-11: Performance results for self-cleaning, averaging estimator with averaging over neighboring observations determined by orbit matching, a 2000-point observation set, and $(m, r) = (4, 4)$. (a) Henon map; (b) Ikeda map.

improvement even at small input SNRs is surprising in light of the fact that no knowledge of the underlying system dynamics is used[1].

We conclude by mentioning one possible extension of this approach, the use of Wiener-type filtering. For example, given a set of suitably chosen neighboring orbit points, a Wiener-type state estimator for $\hat{\boldsymbol{x}}(n_0)$, the estimate of the state at time n_0, is given by

$$\hat{\boldsymbol{x}}(n_0) = \boldsymbol{avg}(n_0) + \boldsymbol{\Lambda}_{avg}(n_0)\left[\boldsymbol{\Lambda}_{avg}(n_0) + \boldsymbol{R}\right]^{-1}\left(\boldsymbol{y}(n_0) - \boldsymbol{avg}(n_0)\right) \tag{4.12}$$

where $\boldsymbol{avg}(n_0)$ denotes the averaged orbit points for time n_0, $\boldsymbol{\Lambda}_{avg}(n_0)$ denotes the (esti-mated) covarian

$_0$ and $\boldsymbol{R}$ denotes

the covariance matrix of the observation noise.

4.3 MMSE State Estimation: Local Techniques

The previous sections of this chapter dealt with nonrandom state estimation, that is, estimation of an unknown, but nonrandom state vector or orbit segment and the emphasis was on ML state estimation. In contrast, this section and the next deal with state estimation in a Bayesian context, with the unknown state vector at a given time treated as a random

[1]Since reporting this self-cleaning, averaging estimator [76, 77], we discovered that similar work was independently pursued and reported in [40]

vector, and the emphasis is on Minimum-Mean-Squared-Error (MMSE) state estimation. In this section we introduce a local, heuristic MMSE state estimator, and in the next, a global, approximate MMSE state estimator. The extended Kalman filter (EKF) provides the foundation for the local estimator discussed in this section, and variations of the estimator are provided for all three problem scenarios—known system dynamics, unknown system dynamics with availability of a clean reference orbit, unknown system dynamics with no reference orbit. The global estimator introduced in the next section is applicable only to the first two of these scenarios. This latter estimator exploits the ergodicity of dissipative, chaotic systems allowing the computationally intractable integral for the MMSE state estimator given by (3.19) to be replaced by an infinite summation which is easier to approximate.

4.3.1 The Extended Kalman Filter and Smoother

The extended Kalman filter (EKF) is a recursive state estimator that can be used with a large class of nonlinear state-space models. Unlike the Kalman filter which is the optimal MMSE estimator for a restricted set of linear state-space models, the EKF is a heuristic algorithm which in general is not the MMSE state estimator for a given, nonlinear, state-estimation problem. Because of this lack of optimality, one can not determine *a priori* the performance of the extended Kalman filter for a specific problem as is possible with the Kalman filter.

In this subsection, we derive the equations for the extended Kalman filter (EKF) for the general DTS/DTO scenario given by (3.1)–(3.2) and repeated here for reference:

$$\boldsymbol{x}(n+1) = \boldsymbol{f}_n(\boldsymbol{x}(n)) + \boldsymbol{g}_n(\boldsymbol{x}(n))\,\boldsymbol{w}(n) \tag{4.13}$$

$$\boldsymbol{y}(n) = \boldsymbol{h}_n(\boldsymbol{x}(n)) + \boldsymbol{v}(n). \tag{4.14}$$

An underlying assumption in the derivation is that the functions $\boldsymbol{f}_n$, $\boldsymbol{g}_n$, and $\boldsymbol{h}_n$ in (3.1) and (3.2) are sufficiently smooth so that they have Taylor series expansions.

To derive the EKF equations for obtaining the state estimate at time $n+1$, one first expands the functions $\boldsymbol{f}_n$ and $\boldsymbol{g}_n$ in Taylor series about the current state estimate $\hat{\boldsymbol{x}}(n|n)$, and the function $\boldsymbol{h}_n$ in a Taylor series about $\hat{\boldsymbol{x}}(n|n-1)$, where as for the Kalman filter $\hat{\boldsymbol{x}}(n|n-1)$ and $\hat{\boldsymbol{x}}(n|n)$ denote the estimates of $\boldsymbol{x}(n)$ based on the observation sets $Y(0, n-1)$

and $Y(0,n)$, respectively. The state estimate used in the expansion for each function is the most recent estimate available when the function requires evaluation. The Taylor series expansions are given by

$$\boldsymbol{f}_n(\boldsymbol{x}(n)) = \boldsymbol{f}_n(\hat{\boldsymbol{x}}(n|n)) + \boldsymbol{F}_n(\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n|n) + \cdots \tag{4.15}$$

$$\boldsymbol{g}_n(\boldsymbol{x}(n)) = \boldsymbol{G}_n + \cdots \tag{4.16}$$

$$\boldsymbol{h}_n(\boldsymbol{x}(n)) = \boldsymbol{h}_n(\hat{\boldsymbol{x}}(n|n-1)) + \boldsymbol{H}_n(\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n|n-1)) + \cdots, \tag{4.17}$$

where

$$\boldsymbol{F}_n = \boldsymbol{D}\{\boldsymbol{f}_n(\hat{\boldsymbol{x}}(n|n))\} \tag{4.18}$$

$$\boldsymbol{G}_n = \boldsymbol{g}_n(\hat{\boldsymbol{x}}(n|n)) \tag{4.19}$$

$$\boldsymbol{H}_n = \boldsymbol{D}\{\boldsymbol{h}_n(\hat{\boldsymbol{x}}(n|n-1))\}. \tag{4.20}$$

Retaining only those terms explicitly shown in the above expansions yields the following approximations to the DTS/DTO state equation (3.1) and observation equation (3.2):

$$\boldsymbol{x}(n+1) = \boldsymbol{f}_n(\hat{\boldsymbol{x}}(n|n)) + \boldsymbol{F}_n(\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n|n)) + \boldsymbol{G}_n\boldsymbol{w}(n) \tag{4.21}$$

$$= \boldsymbol{F}_n\boldsymbol{x}(n) + \boldsymbol{G}_n\boldsymbol{w}(n) + [\boldsymbol{f}_n(\hat{\boldsymbol{x}}(n|n)) - \boldsymbol{F}_n\hat{\boldsymbol{x}}(n|n)] \tag{4.22}$$

$$\boldsymbol{y}(n) = \boldsymbol{h}_n(\hat{\boldsymbol{x}}(n|n-1)) + \boldsymbol{H}_n(\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n|n-1)) + \boldsymbol{v}(n) \tag{4.23}$$

$$= \boldsymbol{H}_n\boldsymbol{x}(n) + \boldsymbol{v}(n) + [\boldsymbol{h}_n(\hat{\boldsymbol{x}}(n|n-1)) - \boldsymbol{H}_n\hat{\boldsymbol{x}}(n|n-1)]. \tag{4.24}$$

In (4.22) and (4.24), $\boldsymbol{F}_n$ and $\boldsymbol{H}_n$ are matrices, and one can evaluate the bracketed expressions since the values of the quantities in these expressions are known at the time they are needed. As a result, these equations are identical to the state and observation equations used by the Kalman filter, with the addition of deterministic input terms. However, the Kalman filter equations can easily be modified to account for deterministic inputs in the state and observation equations. Thus, the Kalman filter can be applied to the system model given by the above equations, and is in fact the MMSE estimator for this model. The resulting filtering equations, provided in Table 4.1, constitute the *extended Kalman filter* for the DTS/DTO model.

Prediction Step

$$\hat{\boldsymbol{x}}(n+1|n) = \boldsymbol{f}_n(\hat{\boldsymbol{x}}(n|n)) \tag{4.25}$$

$$\boldsymbol{P}(n+1|n) = \boldsymbol{F}_n\boldsymbol{P}(n|n)\boldsymbol{F}_n^T + \boldsymbol{G}_n\boldsymbol{Q}(n)\boldsymbol{G}_n^T \tag{4.26}$$

Measurement Update Step

$$\hat{\boldsymbol{x}}(n+1|n+1) = \hat{\boldsymbol{x}}(n+1|n) + \boldsymbol{K}(n+1)\left[\boldsymbol{y}(n+1) - \boldsymbol{h}_{n+1}(\hat{\boldsymbol{x}}(n+1|n))\right] \tag{4.27}$$

$$\boldsymbol{K}(n+1) = \boldsymbol{P}(n+1|n)\boldsymbol{H}_{n+1}^T\left[\boldsymbol{H}_{n+1}\boldsymbol{P}(n+1|n)\boldsymbol{H}_{n+1}^T + \boldsymbol{R}(n+1)\right]^{-1} \tag{4.28}$$

$$\boldsymbol{P}(n+1|n+1) = \left[\boldsymbol{I}_{\mathcal{N}} - \boldsymbol{K}(n+1)\boldsymbol{H}_{n+1}\right]\boldsymbol{P}(n+1|n) \tag{4.29}$$

Initialization

$$\hat{\boldsymbol{x}}(0|-1) = \boldsymbol{m}_0 \tag{4.30}$$

$$\boldsymbol{P}(0|-1) = \boldsymbol{P}_0. \tag{4.31}$$

Table 4.1: The Extended Kalman Filter (EKF) for the DTS/DTO Scenario

A comparison of the equations provided in Tables 3.1 and 4.1 reveals the similarity between the Kalman filter and extended Kalman filter. However, a fundamental difference between the two is that the Kalman filter is the MMSE state estimator for a linear state-space model, whereas the extended Kalman filter is the MMSE state estimator for a linearized state-space model and a heuristic state estimator for the nonlinear state-space model which gives rise to the linearized model. This analogy suggests that the effectiveness of the EKF depends largely on the accuracy of the linear approximations of the nonlinear functions $\boldsymbol{f}_n$, $\boldsymbol{g}_n$, and $\boldsymbol{h}_n$. If the neglected higher-order terms in the Taylor series expansions of these functions are not negligible, the EKF may perform poorly.

An interesting aspect of the EKF concerns the state estimate equation in the prediction step (4.25). Combining this equation with (4.21) leads to the following sequence of equations:

$$\hat{\boldsymbol{x}}(n+1|n) = E(\boldsymbol{x}(n+1)|Y(0,n)) \tag{4.32}$$

$$= \boldsymbol{f}_n(\hat{\boldsymbol{x}}(n|n)) + \boldsymbol{F}_n\left(E(\boldsymbol{x}(n)|Y(0,n)) - \hat{\boldsymbol{x}}(n|n)\right) + E(\boldsymbol{G}_n\boldsymbol{w}(n)) \tag{4.33}$$

$$= \boldsymbol{f}_n(\hat{\boldsymbol{x}}(n|n)), \tag{4.34}$$

since $E(\boldsymbol{x}(n)|Y(0,n)) = \hat{\boldsymbol{x}}(n|n)$ and $E(\boldsymbol{G}_n\boldsymbol{w}(n)) = \boldsymbol{0}$. This implies

$$E(\boldsymbol{x}(n+1)|Y(0,n)) = E(\boldsymbol{f}(\boldsymbol{x}(n))|Y(0,n)) = \boldsymbol{f}(E(\boldsymbol{x}(n)|Y(0,n))), \tag{4.35}$$

which although true for linear systems is not true in general for nonlinear systems.

Just as one can augment the Kalman filter equations to obtain a fixed-lag, linear smoother, one can augment the extended Kalman filter equations to obtain a nonlinear smoother. In fact, the same state and covariance update equations for the fixed-lag, linear smoother given in Table 3.2 can be used for fixed-lag, nonlinear smoothing, with the appropriate substitution of the parameter matrices $\boldsymbol{F}_n$, $\boldsymbol{G}_n$, and $\boldsymbol{H}_n$ calculated for the EKF. In the following sections, we refer to the resulting nonlinear smoother as the extended Kalman smoother (EKS).

4.3.2 Performance Results with Known System Dynamics

In contrast to the Kalman filter for which one can evaluate the error covariance *a priori*, the EKF and EKS are not optimal and one can only evaluate their performance on a specific problem with Monte Carlo simulations. In this section, we present and interpret experimental performance results obtained with the EKF and EKS on the Henon and Ikeda maps when the system dynamics, i.e., the function $\boldsymbol{f}$, are known. All performance results are for the restricted DTS/DTO model given by (3.5) and (3.6) with the added restrictions that $\boldsymbol{h}$ is the identity operator and $\boldsymbol{R}$, the covariance matrix of the observation noise, is diagonal. The results indicate the extreme sensitivity of the EKF and EKS to the driving-noise covariance matrix $\boldsymbol{Q}_n$ in the filtering and smoothing equations.

When using the EKF and EKS, one implicitly assumes that the initial state $\boldsymbol{x}(0)$ is a Gaussian random vector with mean vector $\boldsymbol{m}_0$ and covariance matrix $\boldsymbol{P}_0$, since as indicated in Table 4.1, these values are used to initialize the filter. However, this is often an inappropriate assumption with dissipative, chaotic systems, with a more appropriate assumption often being that the distribution of the initial state is that of the physical measure on the attractor, for which there generally is not a corresponding PDF. In light of this and to facilitate a performance comparison with the global estimator introduced later in this chapter, we obtained all performance results by using the the initial data point $\boldsymbol{y}(0)$ as the initial updated estimate $\hat{\boldsymbol{x}}(0|0)$ and the noise covariance matrix $\boldsymbol{R}$ as the corresponding

error covariance matrix $\boldsymbol{P}(0|0)$. This initialization strategy is equivalent to assuming that $\boldsymbol{P}(0|-1)$ is a diagonal matrix with infinite values on the diagonal corresponding to complete uncertainty in the initial state. Note that using the initial observation to initialize the filter is only possible when the transformation $\boldsymbol{h}$ is invertible. Although this is a suboptimal way to initialize the filter, the performance results obtained with this initialization method were found to be comparable to those obtained with more standard initialization methods.

Figures 4-12 (a) and (b) show the first set of performance results obtained with the EKS and known system dynamics. The performance curves are parameterized by the number of lags used by the EKS, so that for example a lag of i indicates the results are those for the lagged estimator $\hat{\boldsymbol{x}}(n-i|n)$. By definition, a lag of zero corresponds to the standard EKF. Each of the plotted points represents the average improvement in SNR for the two components of the state vector, obtained by estimating a 2000-point orbit segment. Because

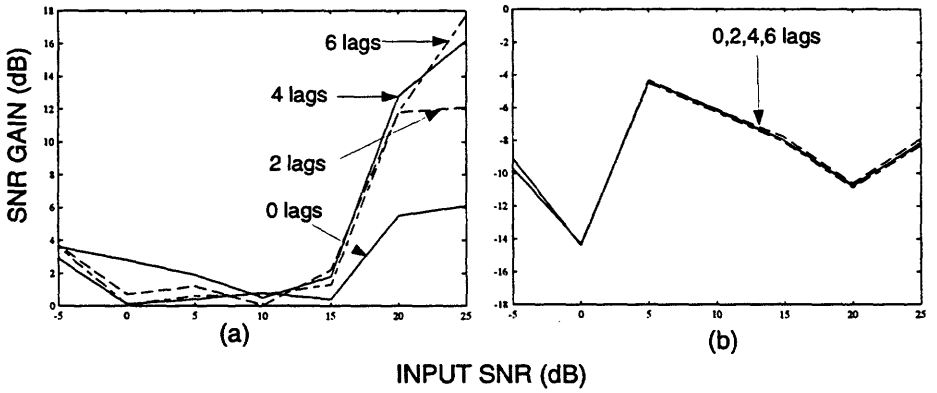

Figure 4-12: Performance results for EKS for different numbers of lags, known system dynamics, and $\mathbf{Q} = [\mathbf{0}]$. (a) Henon map; (b) Ikeda map.

the state equation (3.5) is deterministic and thus there is no driving noise, the driving-noise covariance matrix $\boldsymbol{Q}_n$ was set equal to the zero matrix for all times n in the EKF and EKS equations, so that $\boldsymbol{Q}_n \equiv \boldsymbol{Q} = [\mathbf{0}]$, and the matrix $\boldsymbol{G}_n$ was set equal to the identity matrix $\boldsymbol{I}_{\mathcal{N}}$. The results for the Ikeda map are poor with the negative values indicating that the SNR has actually gotten worse, and the results for the Henon map are mediocre except at larger input SNRs.

Figures 4-13 and 4-14 show the performance results obtained with nonzero, diagonal matrices for $\boldsymbol{Q}_n \equiv \boldsymbol{Q}$ in the EKF and EKS equations, with the constant diagonal value 10^{-5} used for the results in Figures 4-13 (a) and (b) and the value 10^{-3} used for the results in Figures 4-14 (a) and (b). Note that although a nonzero, covariance matrix $\boldsymbol{Q}$ was used

in the EKF and EKS equations, the same deterministic state equation (3.5) was used to generate the data. In other words, the nonzero matrix $\boldsymbol{Q}$ in the EKF and EKS equations corresponds to a fictitious driving noise, or equivalently a conditioning noise used only in the estimator. Overall, the performance results are better than those shown in Figure 4-12;

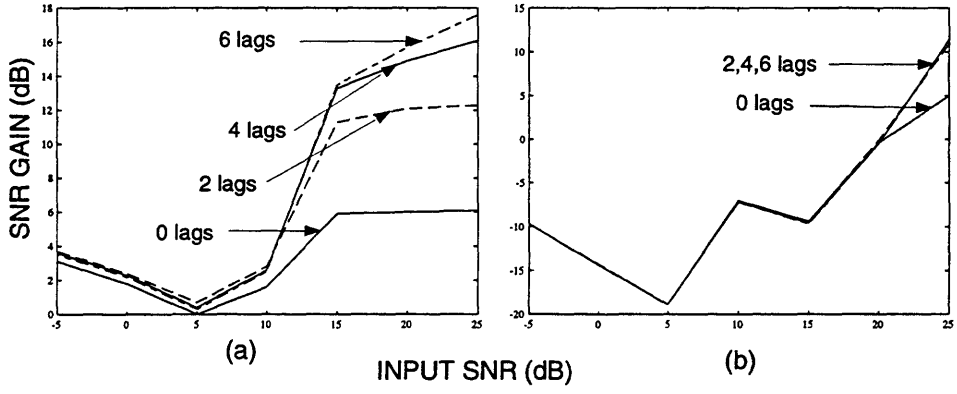

Figure 4-13: Performance results for EKS with known system dynamics, and $\mathbf{Q} = 10^{-5}\ \mathbf{I}_{\mathcal{N}}$. (a) Henon map; (b) Ikeda map.

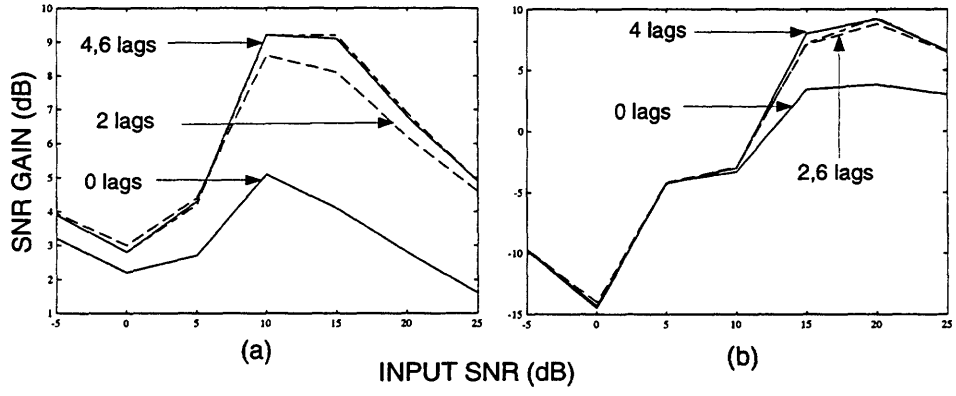

Figure 4-14: Performance results for EKS with known system dynamics, and $\mathbf{Q} = 10^{-3}\ \mathbf{I}_{\mathcal{N}}$. (a) Henon map; (b) Ikeda map.

but especially for the Henon map, concomitant with the performance improvement at some input SNRs is a degradation of performance at other input SNRs. Also, as one might expect, the performance for lagged estimators is considerably better than that for an estimator with no lags, with the improvement apparently saturating at 4 lags.

A comparison of the performance results in Figures 4-12, 4-13, and 4-14 indicates that the EKS is extremely sensitive to the driving-noise covariance matrix $\boldsymbol{Q}$ used in the smoothing equations. One method to circumvent this sensitivity is with careful selection of a unique, driving-noise covariance matrix $\boldsymbol{Q}$ at each SNR. In practical applications, such

tweaking would be undesirable and probably unacceptable, since the SNR may be unknown or time-varying. As such, the performance results suggest that straightforward application of extended Kalman filtering and smoothing techniques to chaotic maps has little practical value. However, as we show in the next section, the combination of extended Kalman smoothing and orbit matching yields a potentially effective state estimator that can be used even if the system dynamics are not known.

4.3.3 Performance Results with Unknown System Dynamics

One can enhance the performance of the EKS using less than full knowledge of the system dynamics. As we show in this section, incorporation of orbit matching in the EKS yields an estimator which does not require full knowledge of the system dynamics, is much less sensitive to the covariance matrix $\boldsymbol{Q}$, and performs better overall than the EKS alone.

As discussed in Section 4.3.1, use of the EKS requires that the system dynamics, i.e., the function $\boldsymbol{f}$ in the state equation, be linearized about the estimated state $\hat{\boldsymbol{x}}(n|n)$ at each time n. In particular, for the restricted DTS/DTO model of interest here the filtering equations use the derivative matrix $\boldsymbol{F}_n = \boldsymbol{D}\{\boldsymbol{f}(\hat{\boldsymbol{x}}(n|n))\}$. When the function $\boldsymbol{f}$ is not known, but a sufficiently long reference orbit is available, one can exploit the topological transitivity of chaotic systems, as was done earlier in the context of ML estimation, to obtain an estimate of the linearized system dynamics at $\hat{\boldsymbol{x}}(n|n)$ needed for extended Kalman filtering and smoothing. In other words, using the reference orbit one can obtain an affine parameter set $\{\boldsymbol{A}_n, \boldsymbol{b}_n\}$, where $\boldsymbol{A}_n$ is an $\mathcal{N} \times \mathcal{N}$-matrix and $\boldsymbol{b}_n$ is an $\mathcal{N}$-element column vector, which for points $\boldsymbol{z}$ in a small neighborhood of $\hat{\boldsymbol{x}}(n|n)$ satisfy the following:

$$\boldsymbol{f}(\boldsymbol{z}) \approx \boldsymbol{A}_n \boldsymbol{z} + \boldsymbol{b}_n. \tag{4.36}$$

The matrix $\boldsymbol{A}_n$ is an estimate of $\boldsymbol{F}_n$, whereas the vector $\boldsymbol{b}_n$ is an offset vector.

One simple, but effective method to estimate the affine parameter set $\{\boldsymbol{A}_n, \boldsymbol{b}_n\}$ is the following. First, find the nearest neighbors (as determined by an appropriate metric) to $\hat{\boldsymbol{x}}(n|n)$ in the reference orbit and the immediate successors to these neighbors, that is, the points in the reference orbit which immediately follow these neighbors. Let $\{\boldsymbol{o}_i\}_{i=1}^{N(n)}$ denote the $N(n)$ orbits points selected as the nearest neighbors to $\hat{\boldsymbol{x}}(n|n)$ and let $\{\boldsymbol{p}_i\}_{i=1}^{N(n)}$ denote the successors to these points, so that $\boldsymbol{p}_i = \boldsymbol{f}(\boldsymbol{o}_i)$. Next, apply least-squares line fitting to

these neighbors and their successors to estimate the affine parameter set. Specifically, find the matrix $\boldsymbol{A}_n$ and $\boldsymbol{b}_n$ which minimize the total prediction error $\mathcal{E}$ given by

$$\mathcal{E} \equiv \sum_{i=1}^{N(n)} \|\boldsymbol{p}_i - (\boldsymbol{A}\,\boldsymbol{o}_i + \boldsymbol{b})\|^2 \tag{4.37}$$

An underlying assumption of this estimation technique is that there is little change in the linearized system dynamics among points in small neighborhoods of $\hat{\boldsymbol{x}}(n|n)$. This assumption is reasonable in light of the underlying restriction in this section that $\boldsymbol{f}$ be differentiable at all points on the chaotic attractor. This method of using a reference orbit to obtain locally linearized estimates of the dynamics of a chaotic system was apparently first proposed in [21], developed further in [22] primarily in the context of prediction, and has been used extensively since for various applications (e.g. [23, 45]).

Having obtained the affine parameter set $\{\boldsymbol{A}_n, \boldsymbol{b}_n\}$, one can perform extended Kalman filtering or smoothing at time n by substituting $\boldsymbol{A}_n$ for $\boldsymbol{F}_n$ in the filtering and smoothing equations and by redefining the equation for the predicted state estimate $\hat{\boldsymbol{x}}(n+1|n)$ given by (4.25) as

$$\hat{\boldsymbol{x}}(n+1|n) = \boldsymbol{A}_n\,\hat{\boldsymbol{x}}(n|n) + \boldsymbol{b}_n, \tag{4.38}$$

In summary, we have the following algorithm for extended Kalman smoothing when the system dynamics are unknown, but a noise-free reference orbit is available.

EKS with Noise-Free Reference Orbit

1. Given the present state estimate $\hat{\boldsymbol{x}}(n|n)$, find the nearest neighbors to $\hat{\boldsymbol{x}}(n|n)$ in the reference orbit. Let $N(n)$ denote the number of selected neighbors.

2. Using the nearest neighbors and their immediate successors in the reference orbit, determine the affine parameter set $(\boldsymbol{A}_n, \boldsymbol{b}_n)$ (where $\boldsymbol{A}_n$ is an $\mathcal{N} \times \mathcal{N}$-matrix and $\boldsymbol{b}_n$ is an $\mathcal{N}$-element column vector) which minimizes the one-step, total squared prediction error among the $N(n)$ chosen neighbors and their successors.

3. Use the affine mapping determined by $(\boldsymbol{A}_n, \boldsymbol{b}_n)$ as an estimate of the system dynamics at $\hat{\boldsymbol{x}}(n|n)$ in the extended Kalman filtering and smoothing equations. In particular, substitute $\boldsymbol{A}_n$ for $\boldsymbol{F}_n$ in these equations, and substitute (4.38) for (4.25) to obtain the predicted estimate $\hat{\boldsymbol{x}}(n+1|n)$.

4. Repeat the procedure until all observations have been processed

There are several parameters associated with the above algorithm such as the number of nearest neighbors used at each time n and the size of the reference orbit. In addition, fundamental elements of the above algorithm have been left unspecified, including the criterion or metric used for selecting nearest neighbors. A logical choice of metric is the Euclidean distance between each point in the reference orbit and the current state estimate. Figures 4-15 and 4-16 show the performance results obtained by using the Euclidean metric for selecting nearest neighbors. The results in each figure were obtained by using a 4000-point

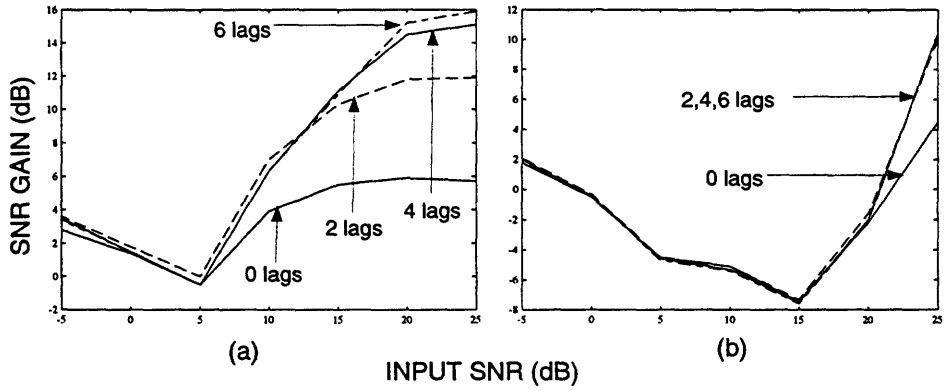

Figure 4-15: Performance results for EKS with 4000-point reference orbit, 15 nearest neighbors, and $\mathbf{Q} = 10^{-5}\,\mathbf{I}_{\mathcal{N}}$. (a) Henon map; (b) Ikeda map.

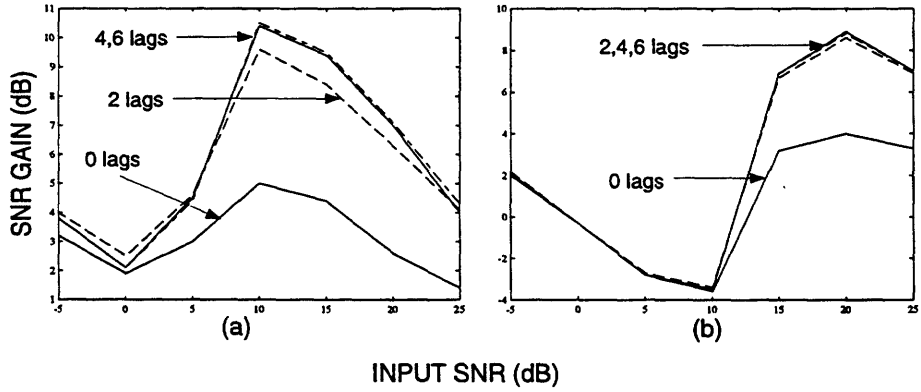

Figure 4-16: Performance results for EKS with 4000-point reference orbit, 15 nearest neighbors, and $\mathbf{Q} = 10^{-3}\,\mathbf{I}_{\mathcal{N}}$. (a) Henon map; (b) Ikeda map.

reference orbit and 15 nearest neighbors for estimating the affine parameters at each time n. In addition, nonzero covariance matrices $\boldsymbol{Q}$ were used in the filtering and smoothing equations, with the same matrices used for the results in Figures 4-15 and 4-16 as were

used for the results in Figures 4-13 and Figures 4-14, respectively. Overall, the results are disappointing (except for the Henon map with larger input SNRs) and comparable to those shown in Figures 4-13 and 4-14 suggesting that the aspects of the system dynamics used by the estimator are available in the reference orbit.

As is the case when the system dynamics are known, one can improve the performance of the EKS when the system dynamics are unknown by carefully selecting, by trial and error, a unique matrix $\boldsymbol{Q}$ for each input SNR. However, one can avoid such undesirable parameter tweaking with the use of an orbit-matching criterion to select nearest neighbors used for estimating each affine parameter set. In particular, one matches, in a weighted least squares sense, the current state estimate $\hat{\boldsymbol{x}}(n|n)$, the previous m state estimates $\hat{\boldsymbol{x}}(n-i|n-i)$, $i = 1, \cdots, m$ and the next r observations $\boldsymbol{y}(n+i)$, $i = 1, \cdots, r$ to orbit segments in the reference orbit and chooses those reference orbit points as nearest neighbors, for which the corresponding orbit segments most closely match. Mathematically, when $\boldsymbol{h}$ is the identity operator, one chooses at time n those reference orbit points $\boldsymbol{o}_i$ for which the following error criterion is smallest:

$$\sum_{j=0}^{m} \left[(\boldsymbol{f}^{-j}(\boldsymbol{o}_i) - \hat{\boldsymbol{x}}(n-j|n-j))^T \boldsymbol{W}^{-1} (\boldsymbol{f}^{-j}(\boldsymbol{o}_i) - \hat{\boldsymbol{x}}(n-j|n-j)) \right] + \sum_{k=1}^{r} \left[(\boldsymbol{f}^{k}(\boldsymbol{o}_i) - \boldsymbol{y}(n+k))^T \boldsymbol{R}^{-1} (\boldsymbol{f}^{k}(\boldsymbol{o}_i) - \boldsymbol{y}(n+k)) \right]. \tag{4.39}$$

If $\boldsymbol{h}$ is not the identity operator, the following criterion is appropriate:

$$\sum_{j=0}^{m} \left[(\boldsymbol{f}^{-j}(\boldsymbol{o}_i) - \hat{\boldsymbol{x}}(n-j|n-j))^T \boldsymbol{W}^{-1} (\boldsymbol{f}^{-j}(\boldsymbol{o}_i) - \hat{\boldsymbol{x}}(n-j|n-j)) \right] + \sum_{k=1}^{r} \left[(\boldsymbol{h}(\boldsymbol{f}^{k}(\boldsymbol{o}_i)) - \boldsymbol{y}(n+k))^T \boldsymbol{R}^{-1} (\boldsymbol{h}(\boldsymbol{f}^{k}(\boldsymbol{o}_i)) - \boldsymbol{y}(n+k)) \right]. \tag{4.40}$$

In the above summations, the term $\boldsymbol{f}^{-j}(\boldsymbol{o}_i)$ denotes the j^{th} reference orbit point before $\boldsymbol{o}_i$, and $\boldsymbol{f}^{k}(\boldsymbol{o}_i)$ denotes the k^{th} reference orbit point after $\boldsymbol{o}_i$. The matrix $\boldsymbol{W}$ is a weighting matrix which one must select. In the examples that follow, all of which use the identity operator for $\boldsymbol{h}$, $\boldsymbol{W}$ was set equal to $\boldsymbol{R}$, even though other choices for $\boldsymbol{W}$ may have been equally or more appropriate.

Figures 4-17 and 4-18 show the performance results obtained with two different choices of the parameter pair (m, r) used in the above criterion for selecting nearest neighbors.

Additional computer experiments have suggested that there is no optimal selection of m

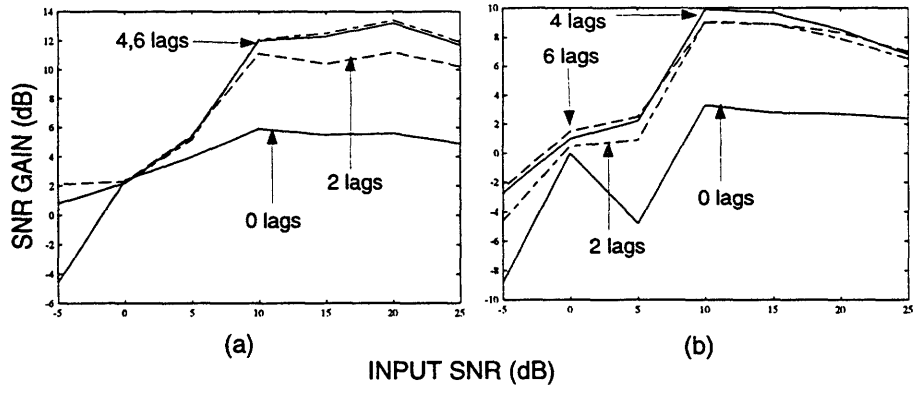

Figure 4-17: Performance results for EKS with 4000-point reference orbit, 15 nearest neighbors, $\mathbf{Q} = 10^{-5}\,\mathbf{I}_{\mathcal{N}}$, and $(m, r) = (3, 3)$. (a) Henon map; (b) Ikeda map.

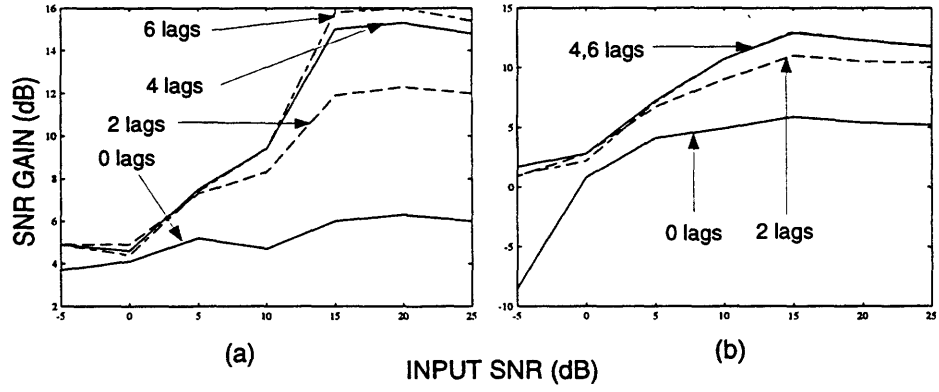

Figure 4-18: Performance results for EKS with 4000-point reference orbit, 15 nearest neighbors, $\mathbf{Q} = 10^{-5}\,\mathbf{I}_{\mathcal{N}}$, and $(m, r) = (1, 4)$. (a) Henon map; (b) Ikeda map.

and r and that various combinations yield comparable results. The plotted results indicate that use of this alternative criterion for selecting nearest neighbors yields an effective state estimator (at least for the chaotic maps chosen), superior to that of the straightforward EKS with known system dynamics.

As one would expect, the size of the reference orbit and the number of nearest neighbors used for estimating the affine parameters strongly influence performance. Figures 4-19 and 4-20 depict the performance results obtained for differently sized reference orbits and different numbers of nearest neighbors, respectively. Both sets of figures show the performance results for a constant estimator lag of 4 and the parameter pair (m, r) equal to $(1, 4)$. The results in Figures 4-19 (a) and (b) indicate performance improves steadily as the size of the reference orbit increases from 500 to 4000 points. The results in Figures 4-20 (a) and (b)

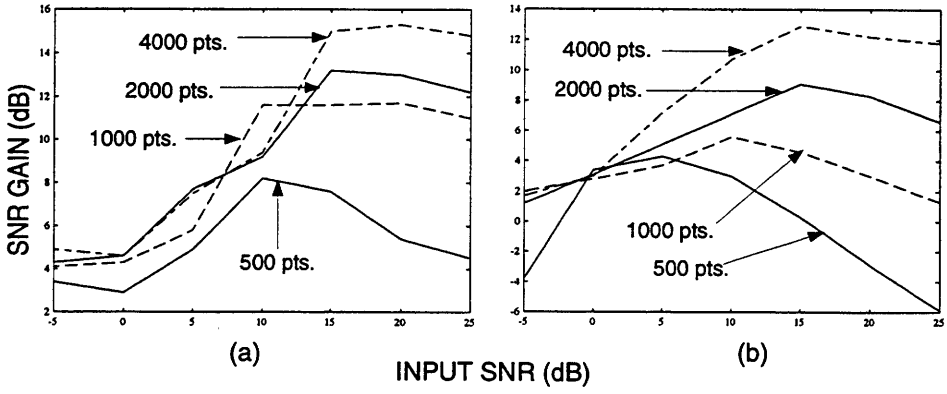

Figure 4-19: Performance results for EKS with 4 lags, 15 nearest neighbors, $\mathbf{Q} = 10^{-5}\,\mathbf{I}_{\mathcal{N}}$, and $(m, r) = (1, 4)$. (a) Henon map; (b) Ikeda map.

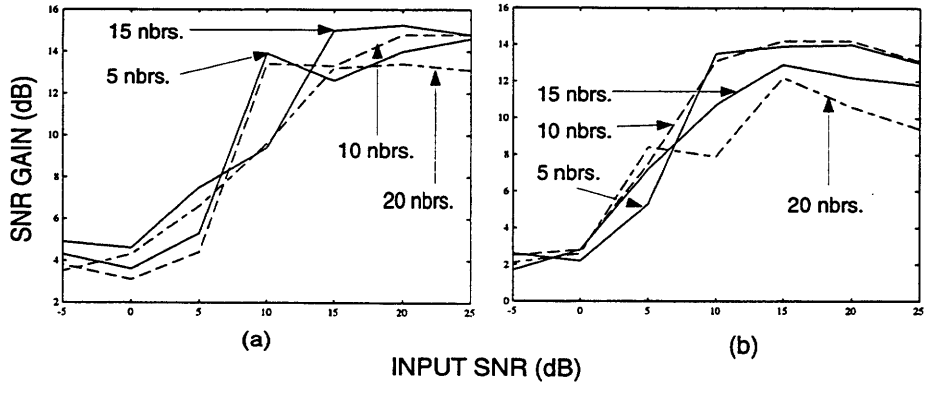

Figure 4-20: Performance results for EKS with 4 lags, 4000-point reference orbit, $\mathbf{Q} = 10^{-5}\,\mathbf{I}_{\mathcal{N}}$, and $(m, r) = (1, 4)$. (a) Henon map; (b) Ikeda map.

indicate that there is no simple relation between neighborhood size and estimator performance, with an apparent degradation in performance as the number of nearest neighbors exceeds some system-dependent threshold.

Surprisingly, this state estimation approach involving affine parameter estimation and extended Kalman smoothing provides nonnegligible performance gain even if no reference orbit is available, the transformation $\boldsymbol{h}$ is the identity operator, and one uses the observation set as the reference orbit in the algorithm outlined above. Figures 4-21 (a) and (b) show the performance results obtained by estimating the first 1000 points of a 2000-point observation set and using the entire observation set as the reference orbit. Surprisingly, although the

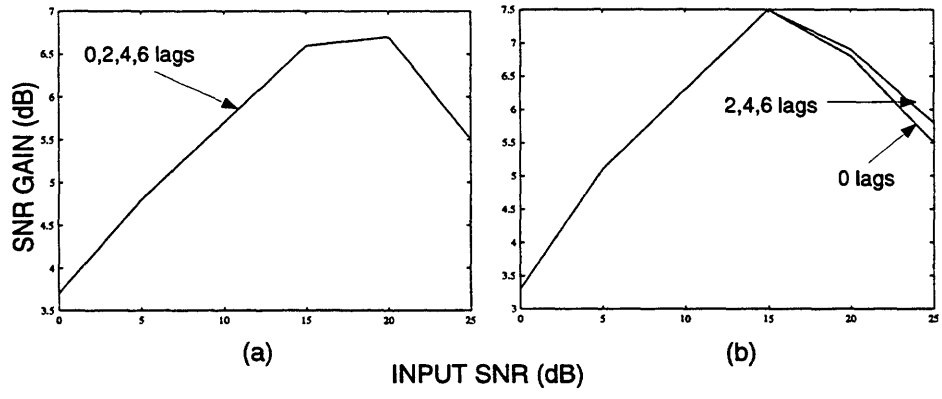

Figure 4-21: Performance results for EKS with no reference orbit, 2000 observations, $\mathbf{Q} = 10^{-4}\,\mathbf{I}_N$, and $(m, r) = (1, 4)$. (a) Henon map; (b) Ikeda map.

performance is not as good as that obtained with a noise-free reference orbit, there is still a nonnegligible SNR gain for each input SNR. Equally surprising is that the SNR gain is nearly the same for different numbers of lags. The plotted SNR gain for each input SNR is not the best achievable; larger gains are obtainable if one carefully selects the noise covariance matrix $\mathbf{Q}$ and parameter pair (m, r) at each input SNR level.

When no reference orbit is available, the question arises as to the additional improvement in performance, if any, obtained by iterating the estimator on the observation set. In other words, does performance continue to improve if one iterates the estimator, using the state estimates at each iteration as the observation set for the next iteration? Computer experiments have suggested that an additional 2-4 dB SNR gain occurs with one additional iteration, but there is little additional gain with more iterations.

A relevant consideration discussed in [27] is that the use of the least-squares method of affine parameter estimation for linear regression problems is applicable only when the

dependent variables and not the independent variables in the regression equations are noise-corrupted. For the affine parameter estimation problem of interest here, both the dependent and independent variables come from the reference orbit. With a noise-free reference orbit, both sets of variables are noise-free; but, with no reference orbit and use of the observation set as the reference orbit, both sets of variables are noise-corrupted. Consequently, the least-squares method used to estimate the affine parameters for the EKS when the observation set is also used as the reference orbit is strictly speaking inappropriate, with a *measurement-error-model* or equivalently *total-least-squares* method [27, 31, 32, 81] theoretically more appropriate. Such an approach was used for the results we presented in [77] in the context of fixed-interval smoothing. However, more recent computer experiments indicated that use of a measurement-error-model method for estimating the affine parameters offered little if any overall performance improvement over use of the simpler least-squares method, when orbit matching was used for selecting nearest neighbors. That is, use of the measurement-error-model method provided performance improvement at some input SNRs and performance degradation at other input SNRs. More importantly, the experiments indicated that the measurement-error-model method was much more sensitive than the least-squares method to filter parameters such as $\boldsymbol{Q}$ and the observation set size. Greater sensitivity to the observation set size is not surprising, as the effectiveness of measurement-error-model methods for parameter estimation often depends strongly on the size of the data set used for estimating the parameters. It was due to the greater sensitivity of the measurement-error-model method than the simpler least-squares method and comparable performance results achieved with both methods (with no parameter fine tuning) that the latter method was used for the above examples.

4.4 MMSE State Estimation: Global Techniques

The state estimators discussed thus far all require that a search be performed for nearest neighbors. In this section, we introduce a global, approximate MMSE state estimator which avoids this computationally intensive requirement. The estimator is applicable when the initial condition $\boldsymbol{x}(0)$ is a random vector with distribution given by the physical measure on the attractor; and as with the state estimators discussed earlier, the estimator uses a noise-free reference orbit.

As discussed in Chapter 3, for a given observation set Y, the MMSE state estimator for $\boldsymbol{x}(n)$ is the conditional mean which is given by (3.16) if the *a posteriori* density $p(\boldsymbol{x}(n)|Y)$ exists and by (3.19) otherwise, with the latter equation repeated here for reference:

$$E(\boldsymbol{x}(n)|Y) = \frac{\int \boldsymbol{x}(n)\, p(Y|\boldsymbol{x}(n))\, d\mu_{\boldsymbol{x}(n)}}{\int p(Y|\boldsymbol{x}(n))\, d\mu_{\boldsymbol{x}(n)}}, \tag{4.41}$$

where $\mu_{\boldsymbol{x}(n)}$ denotes the measure corresponding to the distribution of the state $\boldsymbol{x}(n)$ at time n, and the integration over $\boldsymbol{x}(n)$ is defined in the Lebesgue sense. We know from Chapter 2 that the physical measure on the attractor is an ergodic measure, so that if the initial state is distributed according to this measure, the state at all future times will be distributed according to this measure as well. In light of this and ergodicity, if the distribution of the initial state $\boldsymbol{x}(0)$ is given by the physical measure, then the following holds for the state $\boldsymbol{x}(n)$ at each time n for almost all points $\boldsymbol{z}$ on the chaotic attractor:

$$\frac{\int \boldsymbol{x}(n)\, p(Y|\boldsymbol{x}(n))\, d\mu_{\boldsymbol{x}(n)}}{\int p(Y|\boldsymbol{x}(n))\, d\mu_{\boldsymbol{x}(n)}} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \frac{\boldsymbol{f}^i(\boldsymbol{z})\, p_n(Y|\boldsymbol{f}^i(\boldsymbol{z}))}{\lim_{M\to\infty} \frac{1}{M} \sum_{j=0}^{M-1} p_n(Y|\boldsymbol{f}^j(\boldsymbol{z}))}. \tag{4.42}$$

where $p_n(Y|\boldsymbol{f}^j(\boldsymbol{z}))$ denotes the PDF of the observation set conditioned on $\boldsymbol{f}^j(\boldsymbol{z})$ being the value of the state at time n. Thus, the conditional mean is simply a weighted average of points on a chaotic orbit, where the weight for a point is given by the value of the likelihood function conditioned on that point.

The above summation is not useful for practical MMSE state estimation for two reasons. First, its evaluation requires that an infinite number of terms be calculated. Second, as noted in Section 4.2, the likelihood function rapidly becomes impulse-like as the size of the observation set Y increases. Thus, if one were to attempt to approximate the above expression by summing over a finite number of terms, the value of the likelihood function for each term would for all practical purposes be zero if Y contained more than a few observations.

However, we can use a practical approximation to (4.42) which yields a potentially effective state estimator. First, we approximate the infinite sums in the numerator and denominator with finite sums involving the same number of terms, i.e., $M = N$. Second, we approximate the likelihood function $p_n(Y|\boldsymbol{f}^i(\boldsymbol{z}))$ with $p_n(Y(n-m, n+r)|\boldsymbol{f}^i(\boldsymbol{z}))$, i.e., we only use a subset of observations occurring at times near the time of interest. If $r = 0$ the

estimator is a state filter, whereas if $r > 0$ the estimator is a state smoother. Combining these two approximations yields the following global state estimator $\hat{x}(n)$ for the state at time n:

$$\hat{x}(n) = \sum_{i=0}^{N-1} \frac{f^i(z)\, p_n(Y(n-m, n+r)|f^i(z))}{\sum_{j=0}^{N-1} p_n(Y(n-m, n+r)|f^j(z))} \tag{4.43}$$

where z is an arbitrarily selected point on the chaotic attractor (with the possible exception of a set of points of measure zero). For the results reported here, the set $\{f^i(z)\}_{i=0}^{N-1}$ corresponds to a reference orbit segment.

One property of the estimator that arises because of its use of a rectangular window of observations is that the observation set used at time $n+1$ differs from the observation set used at time n by 2 points. Furthermore, the following relation holds:

$$p_{n+1}(Y(n-m, n+r)|f^i(z)) = p_n(Y(n-m, n+r)|f^{i-1}(z)). \tag{4.44}$$

Therefore, if the reference orbit is allowed to grow by one point at each time n, then one can reduce the computational burden by deriving the likelihoods $p_{n+1}(Y(n-m+1, n+r+1)|f^i(z))$ at time $n+1$ from the likelihoods at time n. Equivalently, if the N likelihoods $\{p_n(Y(n-m, n+r)|f^i(z))\}_{i=0}^{N-1}$ are used in (4.43) at time n, then one can achieve a computational savings at time $n+1$ by using the likelihoods $\{p_{n+1}(Y(n-m+1, n+r+1)|f^i(z))\}_{i=1}^{N}$.

An alternative to the use of a fixed-size, rectangular window of observations at each time n is the use of a growing window of observations with an exponential weighting of past observations, as is often done in recursive filtering applications.

Figures 4-22 and 4-23 depict the performance results obtained with this approach on the Henon and Ikeda maps. In both sets of figures, the curves are parameterized by the pair (m, r) denoting the number of past and future observations used in the likelihood function at each time. Also, for both set of figures, N was set equal to 4000 in (4.43). In Figures 4-22 (a) and (b), $r = 0$; thus, only the present and past observations are used in the likelihood functions resulting in filtered state estimates. In contrast, in Figures 4-23 (a) and (b), $r \neq 0$; thus, future observations are used in the likelihood functions resulting in smoothed state estimates. A comparison of the figures reveals superior performance with smoothing over filtering. A striking feature of the smoothed results is the considerable performance gain even with small input SNRs.

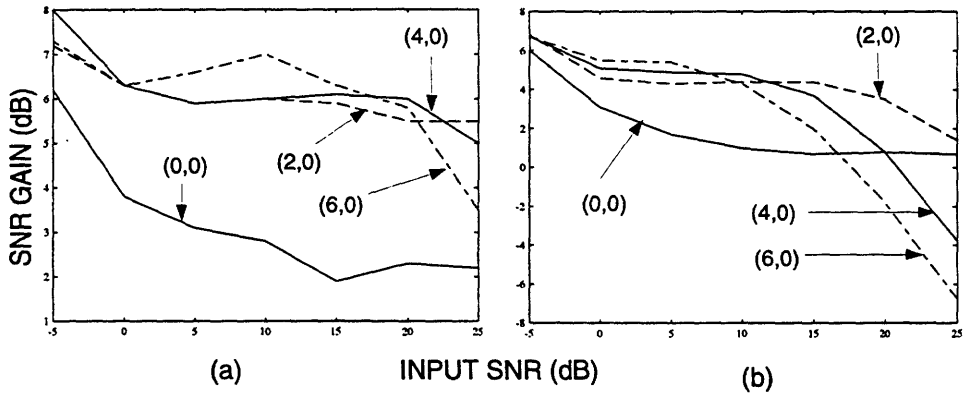

Figure 4-22: Performance results for global estimator with $N = 4000$ and different parameter pairs (m, r). (a) Henon map; (b) Ikeda map.

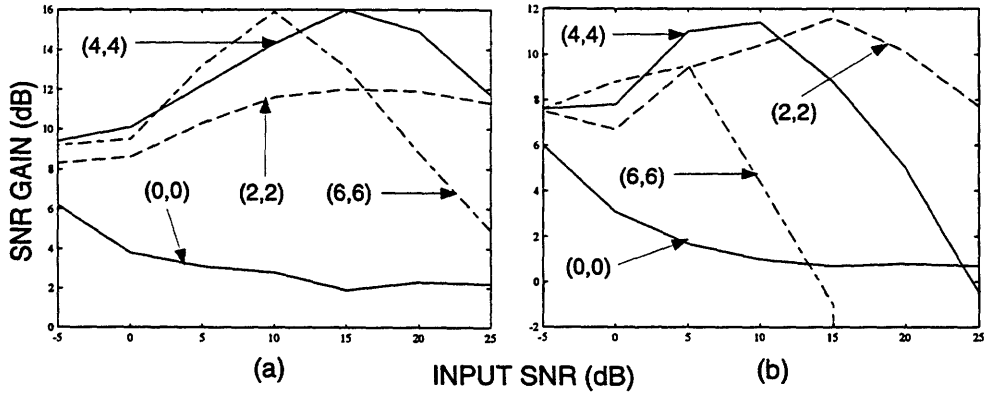

Figure 4-23: Performance results for global estimator with $N = 4000$ and different parameter pairs (m, r). (a) Henon map; (b) Ikeda map.

Figures 4-24 (a) and (b) show the performance results obtained with fixed m and r and different values of N. As might be expected, performance improves as N increases, especially at larger input SNRs.

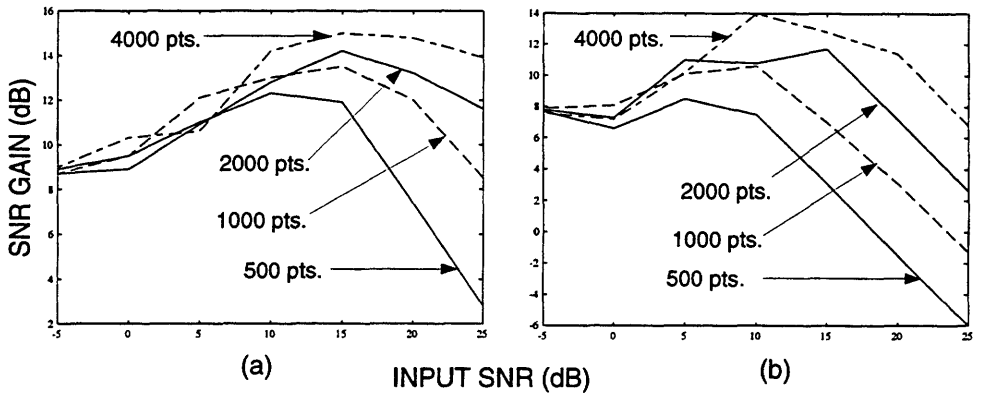

Figure 4-24: Performance results for global estimator with $(m, r) = (3, 3)$ and different values for N. (a) Henon map; (b) Ikeda map.

4.5 Comparison of Estimators

The question arises as to which of the three state estimators introduced in this chapter—the approximate ML, the EKS, or the approximate global MMSE—is the best estimator. In general, it is inappropriate to compare estimators for nonrandom parameters with estimators for random parameters, as the underlying problem scenarios are fundamentally different for the two types of scenarios. However, in light of the underlying assumptions on the *a priori* state distribution used in this chapter, namely that is given by the physical measure on the attractor, and in light of the heuristic nature of the estimators, such a comparison is at least partially justified here.

A comparison of the performance results for the three state estimators suggests that the ML and global MMSE estimators perform comparably, although the experimental results with the MMSE estimator are generally more consistent. In addition, both estimators considerably outperform the EKS with smaller input SNRs; but, the EKS is the superior estimator with larger input SNRs. The poor performance of the EKS with smaller input SNRs and markedly better performance with larger input SNRs is typical performance for a local, MMSE state estimator.

As to the question which estimator is the best, there is no simple answer. Two appealing

aspects of the global MMSE estimator are its ease of implementation as well as its consistent performance when applied to both the Henon and Ikeda maps. Furthermore, performance with larger input SNRs is limited only by the number of terms N used in (4.42). Additional experiments with larger input SNRs have suggested that performance continues to improve as N increases beyond 4000.

One appealing aspect of the EKS is its potential value for self-cleaning, as suggested by the performance results depicted in Figure 4-21. Another appealing aspect is that its performance continues to improve as the input SNR increases beyond 20 dB. However, two unappealing aspects of the EKS are its mediocre performance with extremely small input SNRs and the large number of parameters one must specify when using it. Nonetheless, the performance results presented in this chapter suggest that both the EKS and the global MMSE estimator are potentially effective state estimators with chaotic systems.

Chapter 5

Bounds on State Estimator Performance

5.1 Introduction

As noted in Chapters 3 and 4, practical state estimators for nonlinear systems are often heuristic, and Monte Carlo simulation is needed to assess their performance. When attempting to design and refine state estimators for a given estimation problem, one often has no way of knowing if poor performance of a state estimator is due to the estimator or to a fundamental aspect of the problem itself. As such, it is often useful and desirable to know the best performance achievable by any state estimator for a given estimation problem, or equivalently to have upper bounds on achievable state estimator performance. Ideally, these bounds should be "tight" in the sense that one could derive, at least in theory, an estimator with performance achieving the bounds. However, just as deriving practical, optimal state estimators is often an elusive goal, assessing the tightness of a given performance bound is often an elusive goal as well.

In this chapter, we present and analyze computer simulations of several bounds on the performance of state estimators for chaotic systems. The simulations and analysis indicate that the Lyapunov exponents of chaotic systems strongly affect the achievable performance of state estimators for these systems, and that for a dissipative, chaotic diffeomorphism, there is a positive lower bound on the achievable total error variance when estimating the state at a given time n_0 using observations only for times before n_0 or only for times after

n_0. The simulations suggest that the behavior of the Cramer-Rao bound on state estimator performance for a dissipative, chaotic diffeomorphism is similar to that for an unstable, linear, time-invariant (LTI) system for which the eigenvalues of the state transition matrix are given by the the exponentials of the Lyapunov exponents of the chaotic system. This result may not be surprising, since the definitions of both the Cramer-Rao bound and the Lyapunov exponents involve linearizations of the system dynamics. However, the simulations also reveal that this result holds only when the unknown state being estimated is treated as a nonrandom parameter vector, and that the Cramer-Rao bound and generalizations of this bound on state estimator performance when the unknown state is treated as a random vector with known *a priori* PDF may provide little if any useful information for dissipative, chaotic systems. The simulations also reveal the weakness of the Cramer-Rao bound for noninvertible, chaotic systems, such as the ones considered in Chapter 6, even at moderate input SNRs, and they suggest the value of a generalization of the Cramer-Rao bound known as the Barankin bound for these systems.

The bounds we use in this chapter are not new, and they have been used in the past by others for parameter estimation problems. However, to the best of our knowledge, they have never been used for the types of deterministic systems of interest in this thesis. In addition, the close relation between the behavior of these bounds for multidimensional, nonlinear systems and the Lyapunov exponents of these systems has apparently not been explored in the past, at least not for the problem scenario focused on in this chapter involving small to moderate input SNRs, nonlinear, deterministic system dynamics, and unknown, nonrandom state vectors.

We focus on the state estimation scenario involving unknown, nonrandom state vectors in part because of the undesirable behavior of the performance bounds for random state vectors, as illustrated in the final section of the chapter. As a consequence, many of the results presented in this chapter strictly apply only to unbiased state estimators. The challenging task of possibly extending the results to biased state estimators remains a topic for future research, since it entails the discovery of new error bounds applicable to random parameter vectors.

The next section briefly discusses the estimation problem of interest in this chapter and the performance measures we week to bound. The section also briefly reviews the two general performance bounds—the Cramer-Rao bound and the Barankin bound—emphasized

in the chapter. The derivation of the specific form of these bounds for the estimation problem of interest in this chapter is provided in Appendix A. Section 5.3 provides computer simulations of these bounds for various chaotic systems. In particular, Subsection 5.3.1 presents and qualitatively analyzes simulations of the bounds for two, dissipative, chaotic diffeomorphisms: the Henon map and time-sampled Lorenz flow. The simulations reveal the fundamental influence of system Lyapunov exponents and attractor boundedness on achievable state estimator performance with these systems. Subsection 5.3.2 reveals the weakness of the Cramer-Rao bound when applied to noninvertible, chaotic systems even with moderate input SNRs, and the value of the Barankin bound for use with these systems. In contrast to Sections 5.2 and 5.3 which deal with the problem of estimating nonrandom state vectors, Section 5.4 deals with the problem of estimating random state vectors. Computer simulations of performance bounds applicable to this problem are presented, with the simulations suggesting the limited value of these bounds for state estimation involving dissipative, chaotic systems and the need for novel performance bounds.

Finally, we emphasize that the purpose of this chapter is not to evaluate the specific state estimation algorithms introduced in Chapter 4, but instead to examine the limitations imposed by intrinsic aspects of deterministic, chaotic systems on theoretically achievable state estimator performance with these systems.

5.2 Bounds for Nonrandom State Vectors

As in the previous chapter, we focus on the restricted DTS/DTO scenario given by

$$\boldsymbol{x}(n+1) = \boldsymbol{f}(\boldsymbol{x}(n)) \tag{5.1}$$

$$\boldsymbol{y}(n) = \boldsymbol{h}(\boldsymbol{x}(n)) + \boldsymbol{v}(n) \tag{5.2}$$

where $\boldsymbol{x}(n)$ is the $\mathcal{N}$-dimensional state vector, $\boldsymbol{v}(n)$ is a $\mathcal{P}$-dimensional, zero-mean, Gaussian white-noise sequence with covariance matrix $\boldsymbol{R}$ which is independent of the initial state $\boldsymbol{x}(0)$, and $\boldsymbol{h}$ is a memoryless transformation assumed to be differentiable. Although we derive performance bounds for arbitrary, differentiable $\boldsymbol{h}$, we provide simulations only for the special case in which $\boldsymbol{h}$ is the identity matrix. Because the CTS/DTO scenario can be cast in the form of a DTS/DTO scenario, we use the above state-space model to represent both the DTS/DTO and CTS/DTO scenarios in this chapter.

The state estimation problem we focus on in this section is that of estimating $\boldsymbol{x}(n_0)$, the unknown, but nonrandom state or equivalently orbit point at a fixed time n_0, given an observation set $Y(M,N) \equiv \{\boldsymbol{y}(i)\}_{i=M}^{N}$. As in previous chapters, we often omit the explicit dependence of the observation set on M and N and use Y interchangeably with $Y(M,N)$. We also let $\hat{\boldsymbol{x}}(n_0)$ denote an arbitrary estimator for $\boldsymbol{x}(n_0)$ based on Y, and we let $\boldsymbol{x}_{n_0}$ denote the actual, unknown value of $\boldsymbol{x}(n_0)$.

The two, related performance measures we seek to bound are the error covariance matrix and the trace of this matrix, a quantity which equals the sum of the error variances of the components of $\boldsymbol{x}(n_0)$. For a given estimator $\hat{\boldsymbol{x}}(n_0)$ for the nonrandom state vector $\boldsymbol{x}(n_0)$, the error covariance matrix $\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0))$ is given by

$$\begin{aligned} \boldsymbol{P}(\hat{\boldsymbol{x}}(n_0)) &\equiv E_{Y;\boldsymbol{x}_0}\left\{[\hat{\boldsymbol{x}}(n_0)-\boldsymbol{x}_{n_0}][\hat{\boldsymbol{x}}(n_0)-\boldsymbol{x}_{n_0}]^T\right\} - \boldsymbol{B}(\hat{\boldsymbol{x}}(n_0))\,\boldsymbol{B}^T(\hat{\boldsymbol{x}}(n_0)) \quad (5.3) \\ &= \int \left\{[\hat{\boldsymbol{x}}(n_0)-\boldsymbol{x}_{n_0}][\hat{\boldsymbol{x}}(n_0)-\boldsymbol{x}_{n_0}]^T\right\}\, p(Y;\boldsymbol{x}_{n_0})\,dY \\ &\quad -\boldsymbol{B}(\hat{\boldsymbol{x}}(n_0))\,\boldsymbol{B}^T(\hat{\boldsymbol{x}}(n_0)) \quad (5.4) \end{aligned}$$

where $E_{Y;\boldsymbol{x}_0}$ denotes expectation over the observation set Y given that $\boldsymbol{x}(n_0) = \boldsymbol{x}_{n_0}$, $p(Y;\boldsymbol{x}_{n_0})$ is the likelihood function or equivalently the PDF of Y given that $\boldsymbol{x}(n_0) = \boldsymbol{x}_{n_0}$, and $\boldsymbol{B}(\hat{\boldsymbol{x}}(n_0))$ is the bias of $\hat{\boldsymbol{x}}(n_0)$ as given by

$$\begin{aligned} \boldsymbol{B}(\hat{\boldsymbol{x}}(n_0)) &= E_{Y;\boldsymbol{x}_0}\{\hat{\boldsymbol{x}}(n_0)\} - \boldsymbol{x}_{n_0} \quad (5.5) \\ &= \int \hat{\boldsymbol{x}}(n_0)\,p(Y;\boldsymbol{x}_{n_0})\,dY - \boldsymbol{x}_{n_0} \quad (5.6) \end{aligned}$$

In general, performance bounds for nonrandom parameter vectors, including the two bounds considered in this chapter, are applicable only to unbiased estimators, which are those for which $\boldsymbol{B}(\hat{\boldsymbol{x}}(n_0)) = 0$. Although one can adapt these bounds to handle biased estimators, the resulting bounds are (often undesirably) estimator dependent with some function of the estimator bias appearing in the bound. This unbiasedness constraint on the bounds limits their value, since practical estimators including the ML estimator are inherently biased in many estimation problems. However, the performance bounds for unbiased estimators considered in this chapter provide useful insight into the problem of state estimation with chaotic systems, with this insight being potentially relevant to biased estimators as well. In addition, one of the bounds we consider, the Cramer-Rao bound, is

an asymptotically tight bound that is achievable asymptotically with the ML estimator (an asymptotically unbiased estimator), and a limiting case of the other bound, the Barankin bound, is a tight bound with an explicit form of the unbiased estimator achieving the bound provided in [6].

5.2.1 Cramer-Rao Bound

The Cramer-Rao bound is perhaps the most widely used performance bound for parameter estimators. One advantage of the Cramer-Rao bound over other bounds is the relative ease in explicitly deriving it for many estimation problems, including the problems considered here. Furthermore, for estimation problems involving nonrandom parameters with additive noise, the bound is generally asymptotically tight in the sense that as the input SNR goes to infinity, the Cramer-Rao bound is achieved by an estimator, which in fact is the ML estimator. In addition, at any input SNR if the error covariance matrix of some unbiased estimator satisfies this bound with equality, the estimator is also the ML estimator. Throughout the remainder of the chapter, we distinguish between the Cramer-Rao bounds for estimators of nonrandom and random parameters by denoting the bound for nonrandom parameter estimators as simply the *Cramer-Rao bound* and the bound for random parameter estimators as the *random Cramer-Rao bound.*

Use of the Cramer-Rao bound requires that the likelihood function $p(Y;\boldsymbol{x}(n_0))$ exist and satisfy certain regularity constraints, specifically that it be twice differentiable with respect to the parameter $\boldsymbol{x}(n_0)$ at the actual parameter value $\boldsymbol{x}_{n_0}$ and that both derivatives be integrable with respect to Y. For the estimation problem of interest in this chapter, the Cramer-Rao bound on $P(\hat{\boldsymbol{x}}(n_0))$, the error covariance matrix of of $\hat{\boldsymbol{x}}(n_0)$, is given by [86]

$$\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0)) \geq \boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0}), \tag{5.7}$$

where $\boldsymbol{J}(\boldsymbol{x}_{n_0})$, the *Fisher information matrix*, is given by

$$\boldsymbol{J}(\boldsymbol{x}_{n_0}) = E_{Y;\boldsymbol{x}_0}\left\{\boldsymbol{D}^T_{\boldsymbol{x}(n_0)}\left\{\log p(Y;\boldsymbol{x}_{n_0})\right\}\boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\log p(Y;\boldsymbol{x}_{n_0})\right\}\right\} \tag{5.8}$$

and where $\boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\log p(Y;\boldsymbol{x}_{n_0})\right\}$ denotes the derivative of $\log p(Y;\boldsymbol{x}(n_0))$ evaluated at $\boldsymbol{x}_{n_0}$. We use this convention throughout the chapter, letting the subscript of the derivative operator denote the variable of differentiation and indicating the value of the variable at which

the derivative is to be evaluated as an argument of the function being differentiated.

As shown in Appendix A, for the DTS/DTO scenario given by (5.1) and (5.2), the Fisher information matrix reduces to the following:

$$\boldsymbol{J}(\boldsymbol{x}_{n_0}) = \sum_{i=M}^{N} D^T_{\boldsymbol{x}(n_0)}\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}_{n_0}))\}\, \boldsymbol{R}^{-1} D_{\boldsymbol{x}(n_0)}\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}_{n_0}))\}. \tag{5.9}$$

This expression for $\boldsymbol{J}(\boldsymbol{x}_{n_0})$ is closely related to the expressions defining global and local Lyapunov exponents of $\boldsymbol{f}$. In particular, as discussed in Chapter 3 the Lyapunov exponents of $\boldsymbol{f}$ are the natural logarithms of the eigenvalues of the following matrix [21]:

$$\boldsymbol{\Lambda}_{\boldsymbol{x}} = \lim_{i\to\infty} \left\{ D^T_{\boldsymbol{x}}\{\boldsymbol{f}^i(\boldsymbol{x})\}\, D_{\boldsymbol{x}}\{\boldsymbol{f}^i(\boldsymbol{x})\} \right\}^{\frac{1}{2i}}, \tag{5.10}$$

where $\boldsymbol{x}$ is a point on the attractor (except possibly one of those in a set of measure zero), whereas the local Lyapunov exponents are the natural logarithms of the eigenvalues of the following matrix [2]:

$$\left\{ D^T_{\boldsymbol{x}}\{\boldsymbol{f}^I(\boldsymbol{x})\}\, D_{\boldsymbol{x}}\{\boldsymbol{f}^I(\boldsymbol{x})\} \right\}^{\frac{1}{2I}}, \tag{5.11}$$

for some fixed integer I. A comparison of (5.9) and (5.10) reveals that in the special case that both $\boldsymbol{h}$ and $\boldsymbol{R}$ are $\mathcal{N} \times \mathcal{N}$ identity matrices, the Fisher information matrix $\boldsymbol{J}(\boldsymbol{x}(n))$ consists of a sum of partial products of the infinite product of matrices which determines the Lyapunov exponents of $\boldsymbol{f}$. Equivalently, for this special case the Fisher information matrix consists of a sum of matrices which determine local Lyapunov exponents for increasing values of I in (5.11). As a result, the local and global Lyapunov exponents of $\boldsymbol{f}$ strongly influence the behavior of of $\boldsymbol{J}(\boldsymbol{x}(n_0))$ as suggested by the simulations in Section 5.3.

In many applications, one seeks to bound the trace of $\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0))$, which is the sum of the error variances for the components of $\hat{\boldsymbol{x}}(n_0)$ and which we denote the *total error variance.* That is,

$$Tr\{\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0))\} = \sum_{i=1}^{\mathcal{N}} E_{Y;\boldsymbol{x}_0}\left\{ (\hat{x}_i(n_0) - x_i(n_0))^2 \right\} \tag{5.12}$$

where $Tr\{\cdot\}$ denotes the trace of the bracketed matrix and where

$$\boldsymbol{x}(n_0) = [x_1(n_0), x_2(n_0), \cdots, x_{\mathcal{N}}(n_0)]^T \tag{5.13}$$

$$\hat{\boldsymbol{x}}(n_0) = [\hat{x}_1(n_0), \hat{x}_2(n_0), \cdots, \hat{x}_{\mathcal{N}}(n_0)]^T \tag{5.14}$$

From the Cramer-Rao inequality, it follows that

$$Tr\{\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0))\} \geq Tr\{\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})\}. \tag{5.15}$$

Two relevant facts from linear algebra are that the trace of a matrix equals the sum of its eigenvalues and that the eigenvalues of an invertible matrix equal the reciprocals of the eigenvalues of the inverse. Therefore, if $\{\lambda_i\}$ denote the set of eigenvalues of $\boldsymbol{J}(\boldsymbol{x}_{n_0})$, it follows that

$$Tr\{\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0))\} \geq \sum_i \frac{1}{\lambda_i}. \tag{5.16}$$

Thus, the sum of the reciprocals of the eigenvalues of the Fisher information matrix provides a lower bound on the total error variance. Section 5.3 explores via computer simulation the behavior of the eigenvalues of $Tr\{\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})\}$ as a function of the number of observations for two, dissipative, chaotic diffeomorphisms, the Henon map and time-sampled Lorenz flow, and for a chaotic unit-interval map. The simulations reveal a close relation between these eigenvalues and the Lyapunov exponents of the systems, with the qualitative behavior of the eigenvalues for a given chaotic system similar to that of the eigenvalues for a deterministic, LTI system which has a diagonal state transition matrix with diagonal elements given by the exponentials of the Lyapunov exponents of the chaotic system. This similarity in the behavior of eigenvalues of Fisher information matrices for estimation problems involving chaotic and linear systems may not be surprising in light of the fact that the inverse of the Fisher information matrix for a nonlinear estimation problem is the actual error covariance matrix for the linear estimation problem that arises by linearizing the nonlinear problem about the actual parameter value and estimating small perturbations about this value.

5.2.2 Barankin Bound

As noted in the previous subsection, for nonlinear parameter estimation problems the inverse of the Fisher information matrix is the error covariance matrix for an associated linear estimation problem. As a result, for many nonlinear estimation problems, the Cramer-Rao bound is a fairly tight performance bound only with large input SNRs, with the effect of nonlinearities not accounted for by the Cramer-Rao bound becoming important as the SNR falls below a certain threshold. As we show in the next section, for state estimation with chaotic systems, the Cramer-Rao bound can be a weak bound even with moderately large

input SNRs. To establish the threshold SNR at which the Cramer-Rao bound begins to lose its effectiveness, one must generally consider other performance bounds.

The Barankin bound is not a single bound but a general class, perhaps the most general, of lower bounds on the error moments of unbiased estimators for unknown, nonrandom parameters. Included in this set are the Cramer-Rao and Bhattacharyya bounds. In the most general, probabilistic setting, the Barankin bound is defined on a probability space (Ω, β, μ), where $\Omega = \{\omega\}$ is a set, β is a σ-algebra of subsets of Ω, and μ is a probability measure defined on β, with a family of density functions $\{p(\omega;\theta)\}$ (with respect to the measure μ) assumed to exist on this space. The density functions are indexed by the parameter $\theta \in \Theta$, where Θ is a parameter set. As shown in [6], for any real-valued function $\hat{g}(\omega)$ which is measurable on β and unbiased in the following sense

$$\int \hat{g}(\omega)\, p(\omega;\theta)\, d\mu(\omega) = g(\theta), \tag{5.17}$$

the following inequality holds for all finite m, real constants a_i and parameters $\theta_i \in \Theta$, $i = 1, \cdots, m$ such that the region of support of $p(\omega;\theta_0)$ contains that of each $p(\omega;\theta_i)$:

$$E_{\omega;\theta_0}\left\{(\hat{g}(\omega) - g(\theta_0))^2\right\} \equiv \int (\hat{g}(\omega) - g(\theta_0))^2 p(\omega;\theta_0)\, d\mu(\omega) \tag{5.18}$$

$$\geq \frac{\{\sum_{i=1}^{m} a_i[g(\theta_i) - g(\theta_0)]\}^2}{\int[\sum_{i=1}^{m} a_i L(\omega;\theta_i,\theta_0)]^2 p(\omega;\theta_0) d\mu(\omega)} \tag{5.19}$$

where

$$L(\omega;\theta_i,\theta_0) = \frac{p(\omega;\theta_i)}{p(\omega;\theta_0)}. \tag{5.20}$$

The right-hand-side of the above inequality is the Barankin bound, or more precisely an element of the class of Barankin bounds.

With respect to parameter estimation, the probability space (Ω, β, μ) is the observation space, the density functions $p(\omega;\theta_i)$ are likelihood functions with θ_0 denoting the actual value of the parameter θ one seeks to estimate, and the other θ_i, $i = 1, \cdots, m$, denoting other parameter values which are typically referred to as *test points.* A fundamental, theoretical property of the Barankin bound is that it is a tight bound in the sense that for a fixed parameter value θ_0, the least upper bound of the right-hand-side of (5.19) with respect to all finite m (the number of test points), the test points themselves, and the constants a_i, is achievable by an unbiased estimator, an explicit expression for which is given in [6].

However, the estimator with performance achieving the bound often has little practical value, in part because the form of the estimator is a function of the unknown parameter value.

Appendix A provides the equations for a restricted form of the Barankin bound for vector-valued parameters, originally derived in [58, 59], as specialized to the problem scenario of interest in this chapter with the transformation $\boldsymbol{h}$ in (5.2) equal to the identity operator. As shown in the appendix, this restricted form of the Barankin bound is expressible as a sum of two components—one the inverse of the Fisher information matrix, and the other a positive semidefinite matrix which depends upon the test points, the observation noise covariance matrix, and the number of observations. For the problem scenario of interest here, a strong sensitivity of this second component to the test points, observation noise covariance matrix, and observations arises through the matrix $\boldsymbol{B}$ each element of which is given by

$$B_{ij} = \exp\left\{ \sum_{k=M}^{N} [\boldsymbol{f}^{k-n_0}(\boldsymbol{x}_i(n_0)) - \boldsymbol{f}^{k-n_0}(\boldsymbol{x}_{n_0})]\ \boldsymbol{R}^{-1} \right. \\ \left. \times [\boldsymbol{f}^{k-n_0}(\boldsymbol{x}_j(n_0)) - \boldsymbol{f}^{k-n_0}(\boldsymbol{x}_{n_0})] \right\}, \quad i,j = 1,2\cdots,m \tag{5.21}$$

where m is the number of test points, and $\boldsymbol{x}_j(n_0)$ and $\boldsymbol{x}_i(n_0)$ denote test points, i.e., values of the state vector $\boldsymbol{x}(n_0)$ other than $\boldsymbol{x}_{n_0}$. Note that for $i = j$, B_{ij} is the positive exponential of the weighted distance between orbit segments, one containing the actual parameter value and the other containing a test point, with the inverse of the noise covariance matrix providing the weights. Since the term grows exponentially with this weighted distance, it is extremely sensitive to changes in the noise covariance matrix, number of points in the segments, and the choice of test points. Whereas the inverse of $\boldsymbol{B}$ enters the second component of the bound, the influence of this second component on the overall bound is greatest when the weighted distances between orbit segments are smallest. As a consequence, the influence of the second component of this restricted form of the Barankin bound on the overall bound becomes greater as the input SNR decreases. Such behavior of the second component is desirable, since the first component, the inverse of the Fisher information matrix or equivalently the Cramer Rao bound, becomes tighter as the input SNR increases.

5.3 Computer Simulations

In this section, we present and qualitatively analyze computer simulations of the Cramer-Rao and Barankin bounds for three chaotic systems—the Henon map, time-sampled Lorenz flow, and generalized shift map. The first two systems are dissipative diffeomorphisms whereas the third is neither invertible nor dissipative. The simulations indicate that for dissipative, chaotic diffeomorphisms there is a nonzero lower bound on the total error variance for estimators of $\boldsymbol{x}(n_0)$ given observations for only times less than or equal to n_0 or for only times greater than or equal to n_0, but this lower bound asymptotically approaches zero as the number of observations for times both greater than and less than n_0 increases. The simulations also indicate that the Cramer-Rao bound can be a weak bound even with moderately large input SNRs, for the problem of estimating the initial condition of an orbit segment generated by a noninvertible, chaotic map. This weakness suggests the need for other bounds, such as the Barankin bound, to realistically assess achievable state estimator performance with these systems.

Various aspects of the bounds are considered, but the emphasis is on the behavior of the eigenvalues of the bounding matrices (for the dissipative systems) as a function of the input SNR and the number of observations. This emphasis on eigenvalues arises from the close relation between these eigenvalues and the system Lyapunov exponents. As a consequence of this relation, the behavior of the eigenvalues reflects the influence of the system Lyapunov exponents on the performance bound.

The question arises as to how one chooses test points used in the Barankin bound. As noted in Section 5.2, the restricted Barankin bound used in this chapter is expressible as the sum of two components, one the inverse of the Fisher information matrix, and the other a function of the orbit segments corresponding to these test points, the number of observations, and the observation noise covariance matrix. Also noted in that section is the fact that the influence of this second component on the overall bound is greatest when the distance between these orbit segments and that corresponding to the actual parameter value is smallest. With this in mind, we chose the test points for the dissipative systems used for the simulations using the orbit matching approach introduced in Chapter 4. Specifically, a reference orbit was generated and the m reference orbit points chosen as test points, where m is the number of test points, for which the distances between the corresponding orbit

segments and that corresponding to the actual state vector were smallest. For the unit-interval map, the m points with orbits differing from that of the actual initial condition by the least number of points were used as test points.

5.3.1 Simulations with Dissipative, Chaotic Diffeomorphisms

In this section, we provide computer simulations of the Cramer-Rao and restricted Barankin bounds for the Henon map and time-sampled Lorenz flow. The sampling interval for the Lorenz flow used in all simulations was .005 seconds. In addition, with the exception of the results depicted in Figures 5-1—5-6, all results were obtained with normalized systems for which the experimentally obtained signal variance of each component of the state vector was identical. In other words, the components of the state vector were individually scaled so that they all shared the same variance. Among the reasons for the scaling was that it permitted the use of a single observation noise intensity to achieve the same input SNR for each component of the state vector. Second, it prevented the contribution from one component of the state vector to necessarily dominate the total error variance. The original systems, without scaled state vectors, were used for the results shown in Figures 5-1—5-6 to facilitate the comparison of these systems with diagonal, linear systems having the same Lyapunov exponents. Except where stated otherwise, all simulation results were obtained with $\boldsymbol{R}$, the observation noise covariance matrix, given by $\sigma^2 \boldsymbol{I}_{\mathcal{N}}$ where $\boldsymbol{I}_{\mathcal{N}}$ is the $(\mathcal{N} \times \mathcal{N})$-identity matrix, and with $\boldsymbol{h}$, the output transformation, equal to the identity operator.

Figures 5-1—5-3 (a) depict the eigenvalues of the inverse of the Fisher information matrix for the Henon map as a function of the number of past, future, and both past and future observations, respectively, with an arbitrary point on the attractor used as $\boldsymbol{x}_{n_0}$. Numerical problems arise when calculating the eigenvalues of the Fisher information matrix for the Henon map as the number of observations increases. As a result, relatively few numbers of observations could be used for the figures. Figures 5-1—5-3 (b) depict analogous information for the unstable, diagonal, linear system $\boldsymbol{F}_h$ which has the same set of Lyapunov exponents $\{.42, -1.62\}$ as the Henon map and is given by

$$\boldsymbol{F}_h = \begin{bmatrix} e^{.42} & 0 \\ 0 & e^{-1.62} \end{bmatrix}. \tag{5.22}$$

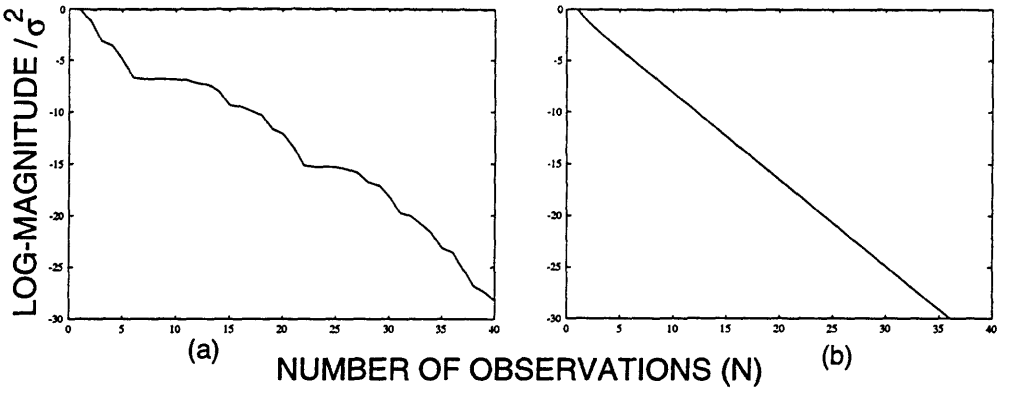

Figure 5-1: Normalized eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ with observation set $Y(n_0, n_0 + N)$. (a) Henon map; (b) $\boldsymbol{F}_h$.

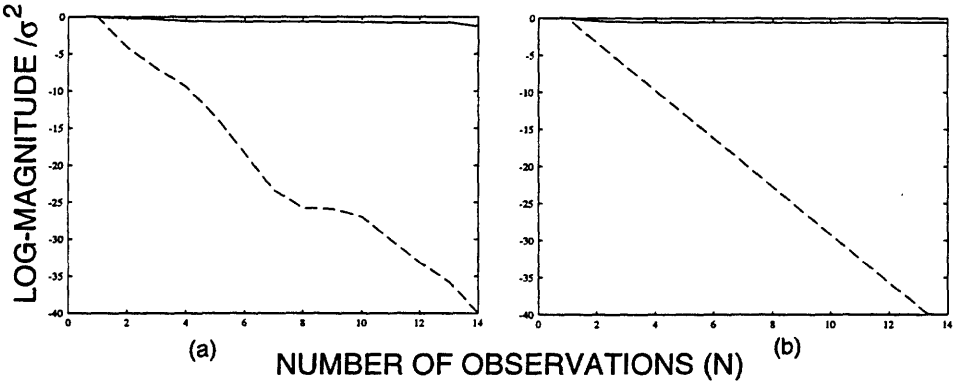

Figure 5-2: Normalized eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ with observation set $Y(n_0 - N, n_0)$. (a) Henon map; (b) $\boldsymbol{F}_h$.

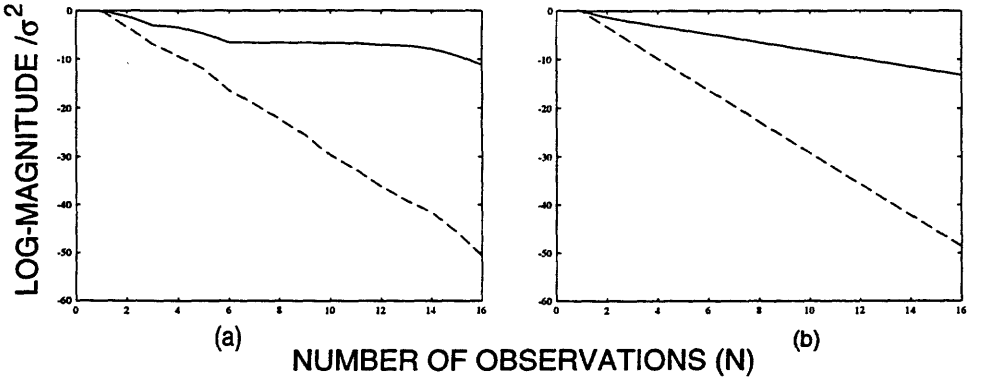

Figure 5-3: Normalized eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ with observation set $Y(n_0 - N, n_0 + N)$. (a) Henon map; (b) $\boldsymbol{F}_h$.

Figures 5-4—5-6 depict analogous information for the time-sampled Lorenz flow and the unstable, diagonal, linear system $\boldsymbol{F}_l$ which has the same set of Lyapunov exponents $\{-.1125, .0075, 0\}$ as the Lorenz flow with sampling interval of .005 seconds and is given by

$$\boldsymbol{F}_l = \begin{bmatrix} e^{.0075} & 0 & 0 \\ 0 & e^{0} & 0 \\ 0 & 0 & e^{-.1125} \end{bmatrix}. \tag{5.23}$$

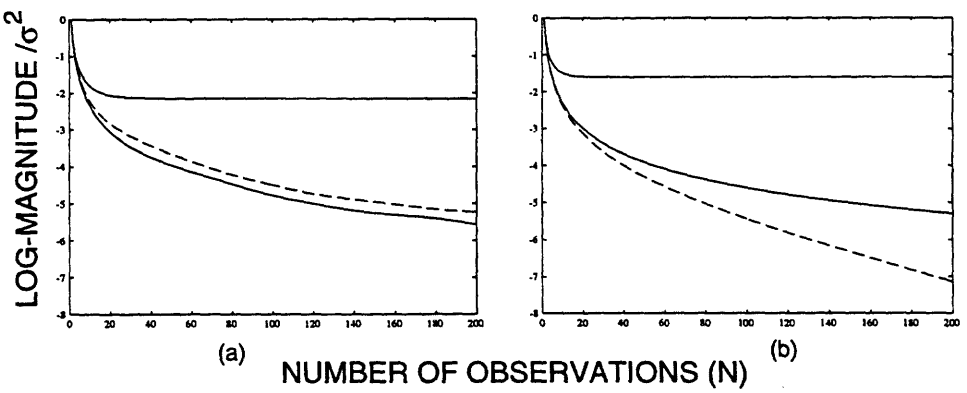

Figure 5-4: Normalized eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ with observation set $Y(n_0, n_0 + N)$. (a) Sampled Lorenz flow; (b) $\boldsymbol{F}_l$.

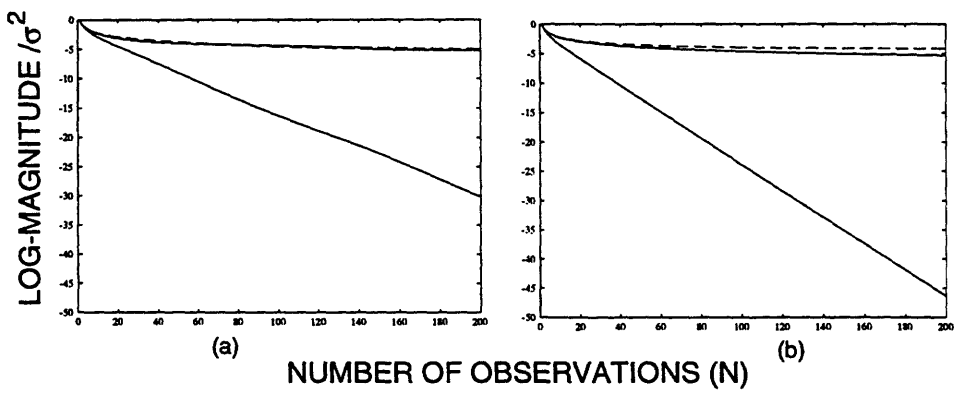

Figure 5-5: Normalized eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ with observation set $Y(n_0 - N, n_0)$. (a) Sampled Lorenz flow; (b) $\boldsymbol{F}_l$.

As indicated by Figures 5-1—5-6, for both the Henon and Lorenz systems the eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ have the same qualitative behavior as those of the corresponding unstable, linear systems. In light of this, it is useful to analyze the inverse of the Fisher information matrix for diagonal, LTI systems to better understand the behavior of the inverse of the

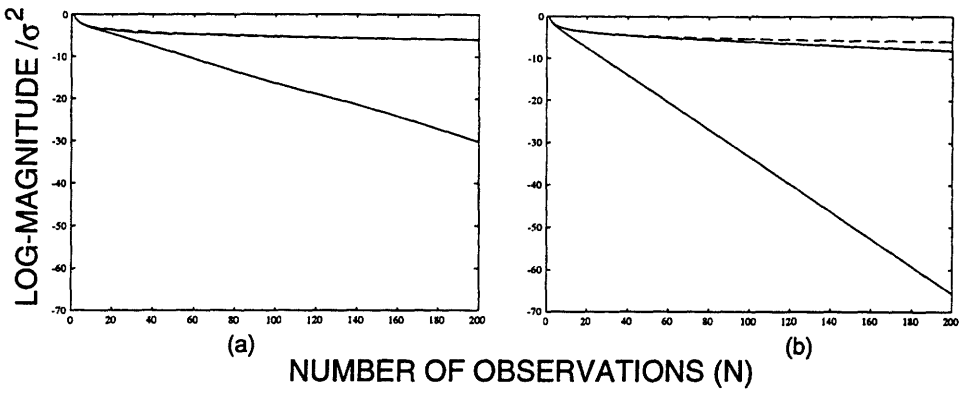

Figure 5-6: Normalized eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ with observation set $Y(n_0 - N, n_0 + N)$. (a) Sampled Lorenz flow; (b) $\boldsymbol{F}_l$.

Fisher information matrix and consequently the Cramer-Rao bound for chaotic systems. Properties of the error covariance matrix of optimal, Bayesian state estimators for arbitrary, noise-driven linear systems have been studied extensively in the past. We only sketch the highlights of these properties relevant to the specific problem of interest here. For an LTI system with state and observation equations given by

$$\boldsymbol{x}(n+1) = \boldsymbol{F}\,\boldsymbol{x}(n) \tag{5.24}$$

$$\boldsymbol{y}(n) = \boldsymbol{x}(n) + \boldsymbol{v}(n) \tag{5.25}$$

were $\{\boldsymbol{v}(n)\}$ is a Gaussian, white-noise sequence with covariance matrix given by $\sigma^2\,\boldsymbol{I}_{\mathcal{N}}$ and where $\boldsymbol{F}$ is an $\mathcal{N} \times \mathcal{N}$ diagonal matrix given by

$$\boldsymbol{F} = \begin{bmatrix} e^{\lambda_1} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_{\mathcal{N}}} \end{bmatrix} \tag{5.26}$$

with the λ_i assumed to be real and distinct, the inverse of the Fisher information matrix

$\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ for arbitrary $\boldsymbol{x}_{n_0}$ and observation set $Y(M,N)$ is given by

$$\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0}) = \sigma^2 \begin{bmatrix} S_1^{-1} & 0 & \cdots & 0 \\ 0 & S_2^{-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & S_N^{-1} \end{bmatrix}, \tag{5.27}$$

where

$$S_j^{-1} = e^{2n_0\lambda_j} \frac{1 - e^{2\lambda_j}}{e^{2M\lambda_j} - e^{(2N+2)\lambda_j}} \tag{5.28}$$

Because $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ is a diagonal matrix, its eigenvalues are given by its diagonal elements, and consequently analyzing the eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ entails analyzing the diagonal elements. Our interest is in the scaling behavior of each diagonal term S_j^{-1} as N becomes increasingly positive and M becomes increasingly negative. The results are the following:

- If $\lambda_j > 0$, then the term $e^{2M\lambda_j}$ grows smaller as M becomes increasingly negative, so that for large negative values of M

$$S_j^{-1} \approx e^{(2n_0-2)\lambda_j}(e^{2\lambda_j} - 1)e^{-2N\lambda_j} \tag{5.29}$$

  which goes to zero only if $N \to \infty$ and does so exponentially at a rate λ_j. Thus, for $\lambda_j > 0$ and a fixed, finite value of N, $S_j^{-1} \geq C(N) > 0$ for some constant $C(N)$ which depends on N.

- If $\lambda_j < 0$, then the situation is reversed, and $e^{2N\lambda_j}$ grows smaller as N becomes increasingly positive, so that for large positive values of N

$$S_j^{-1} \approx e^{2n_0\lambda_j}(1 - e^{2\lambda_j})e^{-2M\lambda_j}, \tag{5.30}$$

  which goes to zero only if $M \to -\infty$ and does so exponentially at a rate λ_j. Thus, for $\lambda_j < 0$ and a fixed, finite value of M, $S_j^{-1} \geq C'(M) > 0$ for some constant $C'(M)$ which depends on M.

- If $\lambda_j = 0$, then

$$S_j^{-1} = \frac{1}{N - M + 1}, \tag{5.31}$$

  which is the reciprocal of the total number of observations. Thus, for $\lambda_j = 0$, S_j^{-1}

scales as the reciprocal of the number of observations.

As noted earlier, for deterministic, LTI systems with real-valued eigenvalues none of which are zero-valued, the Lyapunov exponents are the logarithms of the absolute value of the eigenvalues. Thus, the Lyapunov exponents of $\boldsymbol{F}$ consist of the set $\{\lambda_i\}_{i=1}^{\mathcal{N}}$. As such, the above results indicate that for diagonal, LTI systems with both positive and negative Lyapunov exponents, the bound on the total error variance of estimators for $\boldsymbol{x}_{n_0}$ as given by the sum of the eigenvalues of the Fisher information matrix is nonzero when there are only a finite number of observations for times before n_0, even if the number of observations for times after n_0 goes to infinity. Similarly, the bound is nonzero when there are only a finite number of observations for times after n_0 even if the number of observations for times before n_0 goes to infinity. Only as the number of observations for times both before and after n_0 goes to infinity does the sum of the eigenvalues and thus a lower bound on the total error variance decay asymptotically to zero.

Since dissipative, chaotic systems have both positive and negative Lyapunov exponents and in light of the similar, experimentally observed, qualitative behavior of the eigenvalues of $\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})$ for a chaotic system and the diagonal LTI system with the same Lyapunov exponents, it appears that a similar result holds for chaotic systems, in the sense that there is a nonzero bound on the total error variance of unbiased state estimators for $\boldsymbol{x}_{n_0}$ given only observations for times before or after n_0. The practical implication is that achieving large SNR gains when performing state estimation with dissipative, chaotic systems requires the use of observations both before and after the time of interest.

It is tempting to take this analogy between chaotic systems and unstable LTI systems one step further and conclude that as with deterministic, LTI systems, independent of the input SNR there is an exponential decay in uncertainty, or at least a bound on this uncertainty, in the value of the state at time n_0 as the number of observations for times both greater than and less than n_0 increases, or at worst a decay with rate dominated by the reciprocal of the number of observations. However, as noted earlier the Cramer-Rao bound for nonlinear systems only reflects properties of the locally linearized system. An intrinsic nonlinearity of dissipative, chaotic systems is the boundedness of orbits on the attractor. As one might expect, the size of a chaotic attractor, as given by the variance of orbits on the attractor, influences achievable state estimator performance especially as the attractor size and observation noise intensity become comparable. Whereas the Cramer-Rao bound does

not account for this nonlinearity, the bound becomes weaker as the input SNR decreases, as is the case with most nonlinear systems. The question arises as to the threshold, or equivalently the input SNR, at which the Cramer-Rao bound begins to lose its effectiveness and the rate at which it does so for chaotic systems. One particularly illustrative way to at least qualitatively answer this question is with the use of the Barankin bound.

Figures 5-7—5-12 (a) depict the eigenvalues of the Cramer-Rao bound as functions of the number of past, future, and both past and future observations, respectively with an input SNR of 10 dB and with an arbitrarily selected point on the attractor used as x_{n_0}. In contrast to the earlier figures, the plotted results are not normalized by the noise variance σ^2. Figures 5-7—5-12 (b) depict analogous information for the restricted Barankin bound with 5 test points.

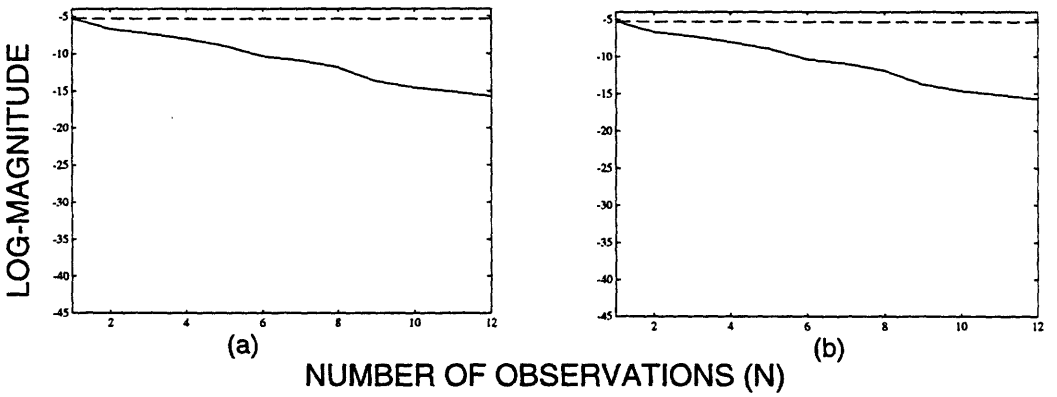

Figure 5-7: Eigenvalues of Cramer-Rao and Barankin bounds with observation set $Y(n_0, n_0 + N)$ for Henon map. (a) Cramer-Rao; (b) Barankin.

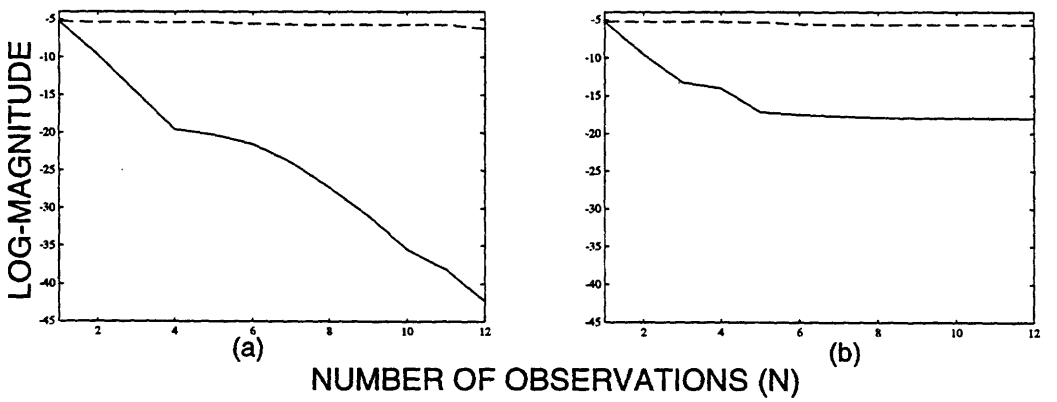

Figure 5-8: Eigenvalues of Cramer-Rao and Barankin bounds with observation set $Y(n_0 - N, n_0)$ for Henon map. (a) Cramer-Rao; (b) Barankin.

A comparison of the results in part (a) with those in part (b) for each of the figures

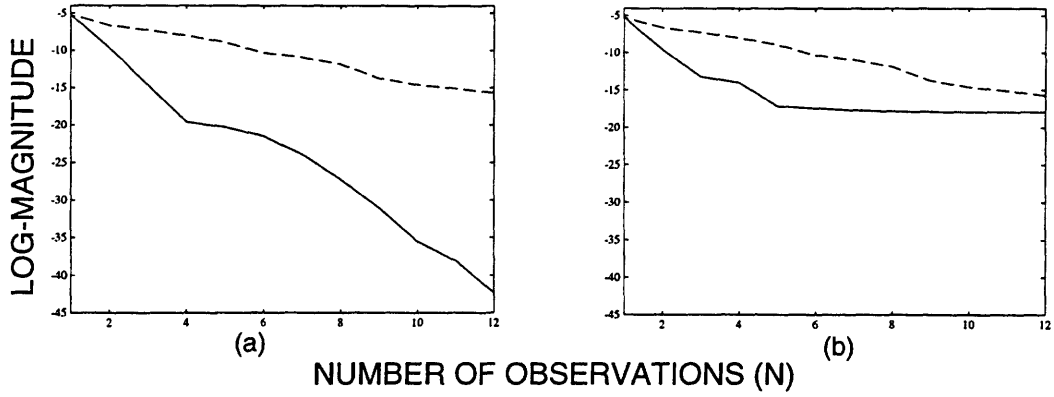

Figure 5-9: Eigenvalues of Cramer-Rao and Barankin bounds with observation set $Y(n_0 - N, n_0 + N)$ for Henon map. (a) Cramer-Rao; (b) Barankin.

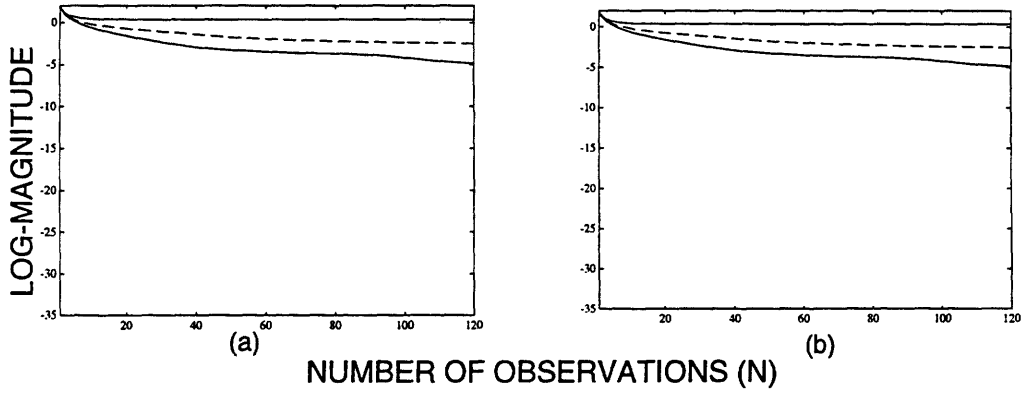

Figure 5-10: Eigenvalues of Cramer-Rao and Barankin bounds with observation set $Y(n_0, n_0 + N)$ for time-sampled Lorenz flow. (a) Cramer-Rao; (b) Barankin.

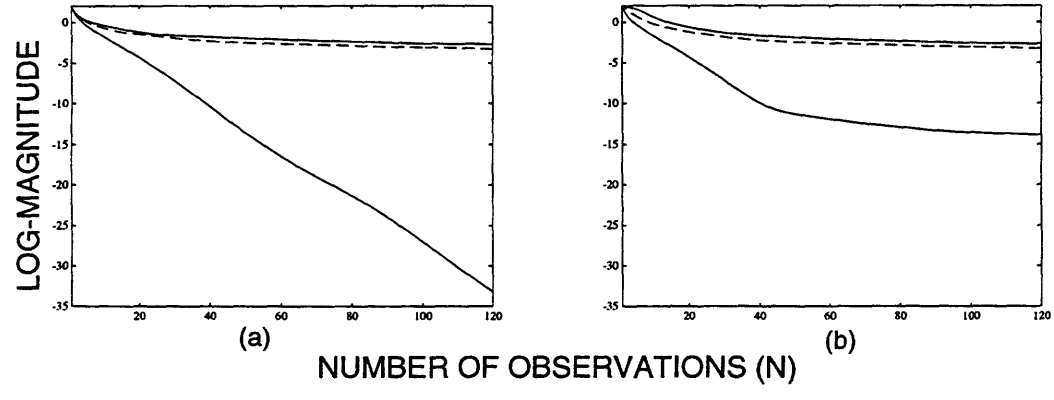

Figure 5-11: Eigenvalues of Cramer-Rao and Barankin bounds with observation set $Y(n_0 - N, n_0)$ for time-sampled Lorenz flow. (a) Cramer-Rao; (b) Barankin.

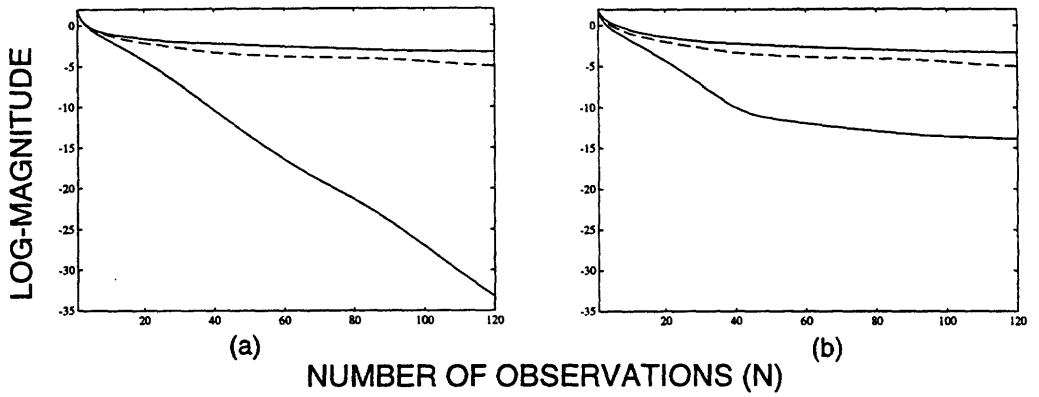

Figure 5-12: Eigenvalues of Cramer-Rao and Barankin bounds with observation set $Y(n_0 - N, n_0 + N)$ for time-sampled Lorenz flow. (a) Cramer-Rao; (b) Barankin.

indicates that the decay rate for the most rapidly decaying eigenvalue with observations at times before n_0 is much smaller for the Barankin bound than for the Cramer-Rao bound, thereby suggesting that the exponential decay rates associated with the Cramer-Rao bound are overly optimistic. Figures 5-13 and 5-14 depict the eigenvalue sums for the Cramer-Rao and Barankin bounds as a function of the input SNR for two different values of N in the observation sets $Y(n_0 - N, n_0 + N)$. The figures indicate that for both systems, the the Cramer-Rao and Barankin bounds on the total error variance deviate at input SNRs below a threshold that depends upon both the system and the number of observations. The results suggest the value of the Barankin bound on assessing achievable state estimation performance with dissipative, chaotic systems and input SNRs smaller than 20 dB.

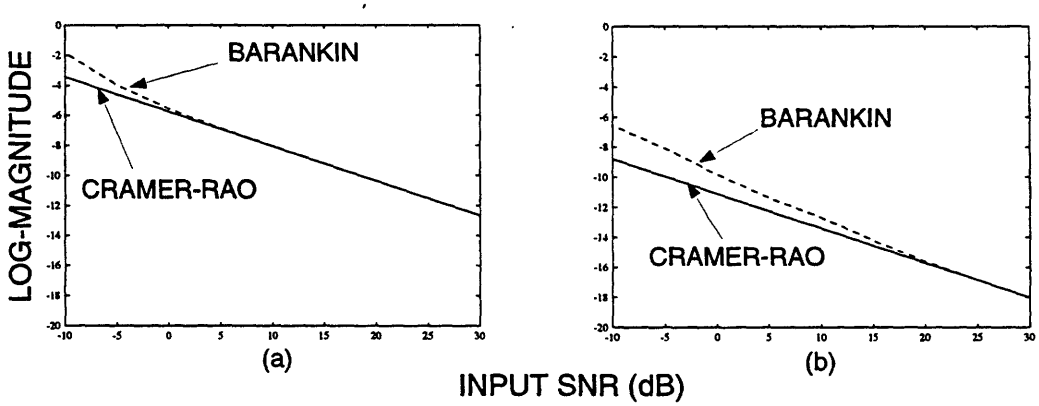

Figure 5-13: Sum of eigenvalues of Cramer-Rao and Barankin bounds with observation set $Y(n_0 - N, n_0 + N)$ for Henon map. (a) $N = 6$; (b) $N = 12$.

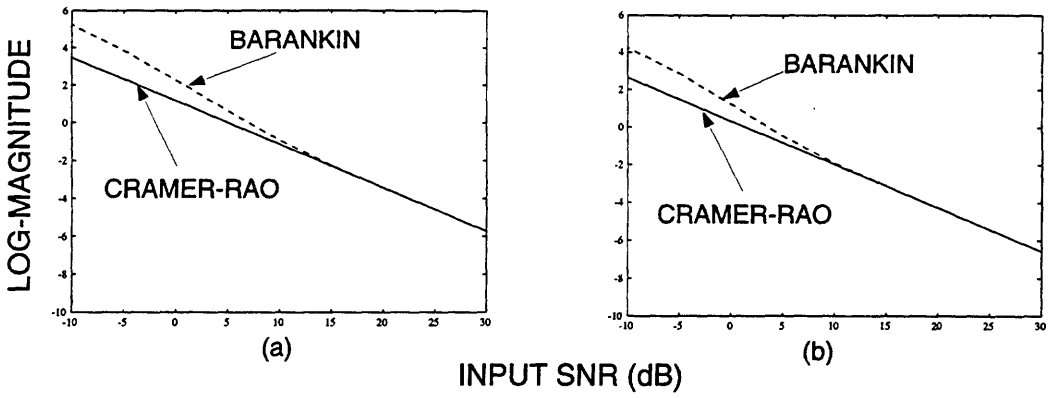

Figure 5-14: Sum of eigenvalues of Cramer-Rao and Barankin bounds with observation set $Y(n_0 - N, n_0 + N)$ for time-sampled Lorenz flow. (a) $N = 20$; (b) $N = 40$.

5.3.2 Simulations with Unit-Interval Maps

In this section, we consider performance bounds on unbiased estimators of the initial condition for a noninvertible, chaotic map. Although we focus on unit-interval maps, the results are relevant to multidimensional chaotic maps as well. In particular, we consider bounds on estimators for the initial condition $x(0)$ of the shift map with observation set $Y(0, N)$, where the general form of the shift map is given by

$$x(n+1) = f(x(n)) = \alpha\, x(n) \pmod{\beta} \tag{5.32}$$

where $\pmod{\beta}$ denotes *modulo* β and where α is an integer with absolute value greater than one. Figure 5-15 depicts the function f for the parameter pair ($\alpha = 4$, $\beta = 1$), the pair used for all examples in this section.

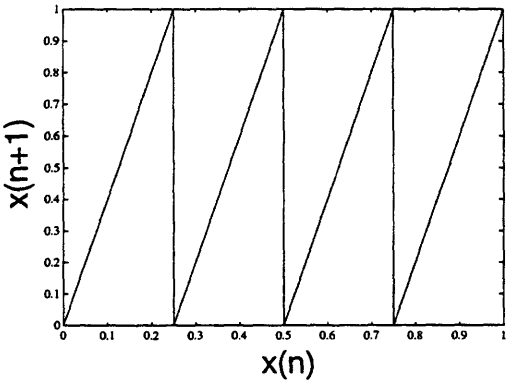

Figure 5-15: Shift map with parameter pair ($\alpha = 4$, $\beta = 1$).

As indicated in the figure, the shift map is discontinuous, with $\alpha - 1$ points of disconti-

nuity. However, the map is differentiable except at these $\alpha - 1$ points and has the constant derivative of α. The Cramer-Rao bound is defined for each initial condition x_0 which is not a discontinuity point of any of the composed functions $\{f^i\}_{i=0}^N$. For each such initial condition and observation set $Y(0, N)$, a straightforward derivation reveals that the inverse of the Fisher information $J^{-1}(x_0)$ is given by

$$J^{-1}(x(0)) = \sigma^2 \frac{\alpha^2 - 1}{\alpha^{2(N+1)} - 1} \tag{5.33}$$

where σ^2 is the variance of the observation noise. This expression decays exponentially with the number of observations at a rate α^{-2}. A similar result, exponential decay of the Cramer-Rao bound with the number of future observations, holds for all other unit interval maps having positive Lyapunov exponents. However, for the shift map, α points are mapped by f to each point; as a consequence and independent of the number of observations, $\alpha - 1$ orbit segments differ from that generated by x_0 by a single point. Similarly, $\alpha^i - 1$ orbit segments differ from that generated by x_0 by at most i points for each integer i. In light of this, one might expect the Cramer-Rao bound to be a weak bound for the shift map even with moderately large input SNRs.

Figures 5-16 and 5-17 confirm this expectation. The figures depict the Cramer-Rao and

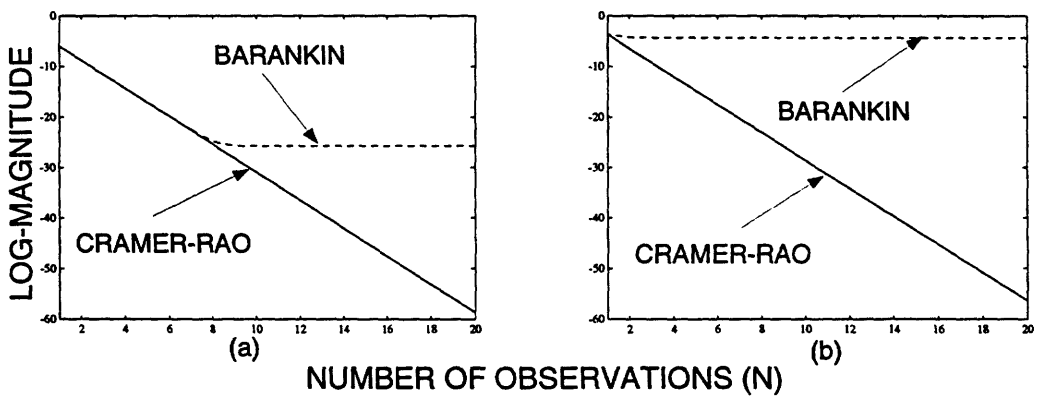

Figure 5-16: Cramer-Rao and Barankin bounds for initial condition estimators of the shift map with parameter pair ($\alpha = 4$, $\beta = 1$) and with observation set $Y(0, N)$. (a) Input SNR = 15 dB; (b) Input SNR = 5 dB.

Barankin bounds on the performance of unbiased estimators for $x(0)$ for the shift map as a function of the number of observations and input SNR, respectively. Eight test points were used in the Barankin bound for the results shown in the figures, with the orbit segments of the test points differing from that of the actual initial condition by at most 2 orbit points.

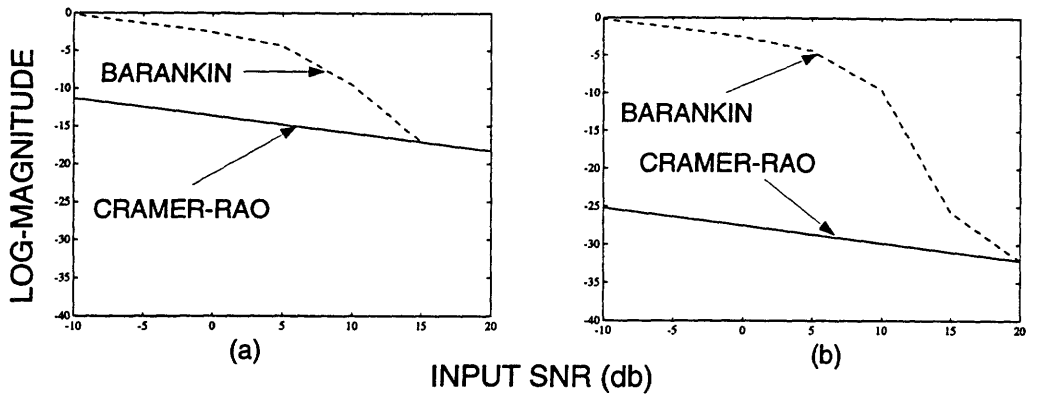

Figure 5-17: Cramer-Rao and Barankin bounds for initial condition estimators of the shift map with parameter pair ($\alpha = 4$, $\beta = 1$) and with observation set $Y(0, N)$. (a) $N = 4$; (b) $N = 8$.

The figures indicate that the Cramer-Rao bound becomes a progressively weaker bound as the number of observations increases. In the figures, the nonzero limiting behavior of the Barankin bound arises from the noninvertibility of the shift map. Divergence of the Cramer-Rao and Barankin bounds is not unique to the shift map, but instead can be expected with any chaotic, unit-interval map and certain, noninvertible, multidimensional chaotic maps as well.

5.4 Bounds for Random State Vectors

A fundamental limitation of performance bounds for estimators of nonrandom parameters is that they are either applicable only to unbiased estimators or they are estimator-dependent with the dependency generally a function of the estimator bias. In contrast, many performance bounds for estimators of random parameters are applicable to both biased and unbiased estimators, and the bounds are not estimator dependent. In this section, we briefly consider performance bounds for state estimators of dissipative, chaotic systems, when the unknown state vector $\boldsymbol{x}(n_0)$ is a random vector with known *a priori* PDF. Experimental results are presented which suggest that widely used performance bounds for estimators of random parameters have limited value with dissipative, chaotic maps. The specific problem we consider is that of estimating the random initial condition $\boldsymbol{x}(0)$ of an orbit segment generated by a dissipative, chaotic diffeomorphism given the $(N+1)$-point observation set $Y(0, N)$ for the special case in which $p(\boldsymbol{x}(0))$, the *a priori* PDF of the initial condition, is Gaussian with mean vector $\boldsymbol{x}_0$ and covariance matrix $\gamma^2 \boldsymbol{I}_N$. All experimental results in the

section were obtained with the time-sampled Lorenz flow. Because the Lorenz attractor has a bounded region of attraction, the mean vector and covariance matrix of $p(\boldsymbol{x}(0))$ used in obtaining the experimental results were chosen to ensure that the probability of an initial condition lying outside this region was extremely small.

5.4.1 Random Cramer-Rao Bound

In general, if the joint probability measure on the initial condition $\boldsymbol{x}(0)$ and observation set Y has a corresponding PDF $p(Y, \boldsymbol{x}(0))$, then the error correlation matrix $\boldsymbol{P}_R(\hat{\boldsymbol{x}}(0))$ for any estimator $\hat{\boldsymbol{x}}(0)$ of $\boldsymbol{x}(0)$ is given by

$$\begin{aligned} \boldsymbol{P}_R(\hat{\boldsymbol{x}}(0)) &\equiv E_{Y,\boldsymbol{x}(0)}\left\{[\hat{\boldsymbol{x}}(0)-\boldsymbol{x}(0)][\hat{\boldsymbol{x}}(0)-\boldsymbol{x}(0)]^T\right\} && (5.34) \\ &= \int\int \left\{[\hat{\boldsymbol{x}}(0)-\boldsymbol{x}(0)][\hat{\boldsymbol{x}}(0)-\boldsymbol{x}(0)]^T\right\} p(Y,\boldsymbol{x}(0))\, dY\, d\boldsymbol{x}(0). && (5.35) \end{aligned}$$

where as indicated above $E_{Y,\boldsymbol{x}(0)}$ denotes expectation over the joint PDF $p(Y, \boldsymbol{x}(0))$. The trace of $\boldsymbol{P}_R(\hat{\boldsymbol{x}}(0))$ yields the total mean-squared (estimation) error (MSE) for the components of $\hat{\boldsymbol{x}}(0)$. The Cramer-Rao bound on $\boldsymbol{P}_R(\hat{\boldsymbol{x}}(0))$, hereafter denoted the *random Cramer-Rao bound*, is given by

$$\boldsymbol{P}_R(\hat{\boldsymbol{x}}(0)) \geq \boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0)) \tag{5.36}$$

where $\boldsymbol{J}_R(\boldsymbol{x}(0))$, the Fisher information matrix for the random parameter vector $\boldsymbol{x}(0)$, hereafter denoted the *random Fisher information matrix*, is given by [86]

$$\boldsymbol{J}_R(\boldsymbol{x}(0)) = E_{Y,\boldsymbol{x}(0)}\left\{D^T_{\boldsymbol{x}(0)}\left\{\log p(Y,\boldsymbol{x}(0))\right\} D_{\boldsymbol{x}(0)}\left\{\log p(Y,\boldsymbol{x}(0))\right\}\right\} \tag{5.37}$$

or equivalently by

$$\begin{aligned} \boldsymbol{J}_R(\boldsymbol{x}(0)) &= E_{Y,\boldsymbol{x}(0)}\left\{D^T_{\boldsymbol{x}(0)}\left\{\log p(Y|\boldsymbol{x}(0))\right\} D_{\boldsymbol{x}(0)}\left\{\log p(Y|\boldsymbol{x}(0))\right\}\right\} \\ &\quad + E_{\boldsymbol{x}(0)}\left\{D^T_{\boldsymbol{x}(0)}\left\{\log p(\boldsymbol{x}(0))\right\} D_{\boldsymbol{x}(0)}\left\{\log p(\boldsymbol{x}(0))\right\}\right\} && (5.38) \\ &= E_{\boldsymbol{x}(0)}\left\{\boldsymbol{J}(\boldsymbol{x}(0))\right\} \\ &\quad + E_{\boldsymbol{x}(0)}\left\{D^T_{\boldsymbol{x}(0)}\left\{\log p(\boldsymbol{x}(0))\right\} D_{\boldsymbol{x}(0)}\left\{\log p(\boldsymbol{x}(0))\right\}\right\} && (5.39) \end{aligned}$$

where $\boldsymbol{J}(\boldsymbol{x}(0))$ has the same form as the Fisher information matrix for nonrandom parameters used in the previous section. Use of the random Cramer-Rao bound requires that $p(Y, \boldsymbol{x}(n_0))$ exist and satisfy certain regularity constraints, specifically that it be twice differentiable with respect to $\boldsymbol{x}(n_0)$ and that both derivatives be integrable with respect to both $\boldsymbol{x}(n_0)$ and Y.

Since $p(\boldsymbol{x}(0))$ is a Gaussian PDF with covariance matrix $\gamma^2 \boldsymbol{I}_{\mathcal{N}}$, it follows that

$$E_{\boldsymbol{x}(0)} \left\{ D^T_{\boldsymbol{x}(0)} \left\{ \log p(\boldsymbol{x}(0)) \right\} D_{\boldsymbol{x}(0)} \left\{ \log p(\boldsymbol{x}(0)) \right\} \right\} = \gamma^{-2} \boldsymbol{I}_{\mathcal{N}}, \tag{5.40}$$

so that for the problem of interest here $\boldsymbol{J}_R(\boldsymbol{x}(0))$ reduces to

$$\boldsymbol{J}_R(\boldsymbol{x}(0)) = E_{\boldsymbol{x}(0)} \left\{ \boldsymbol{J}(\boldsymbol{x}(0)) \right\} + \gamma^{-2} \boldsymbol{I}_{\mathcal{N}}. \tag{5.41}$$

As indicated by (5.41), the first component of $\boldsymbol{J}_R(\boldsymbol{x}(0))$ involves an averaging of matrices, where each matrix has the same form as the Fisher information matrix for a nonrandom parameter. As suggested by the simulation results that follow, because of this averaging the behavior of the random Fisher information matrix for initial condition estimators with dissipative, chaotic maps differs considerably from that of the nonrandom Fisher information matrix (i.e, the Fisher information matrix for nonrandom parameters). The results indicate that this averaging can cause the random Fisher information matrix to exhibit undesirable behavior which severely limits the value of the random Cramer-Rao bound for disspative, chaotic systems.

Figures 5-18 and 5-19 depict the eigenvalues of $\boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0))$ for the time-sampled Lorenz flow as a function of N for input SNRs of 20 dB and 30 dB, respectively. For these results and all the results reported in this section, the expectation over $\boldsymbol{x}(0)$ in (5.41) was performed using Monte Carlo simulation, with $\boldsymbol{J}(\boldsymbol{x}(0))$ calculated for 1000 samples of $\boldsymbol{x}(0)$ randomly selected according to $p(\boldsymbol{x}(0))$ and the resulting matrices averaged. As indicated by both pairs of figures, the behavior of the eigenvalues differs considerably for the different values of γ, with the behavior for the smaller value similar to that observed earlier for nonrandom parameters. For the larger value of γ, all of the eigenvalues decrease with increasing N, with none exhibiting the limiting behavior exhibited by those for the smaller value of γ.

A nonrigorous explanation for this eigenvalue behavior is the following. Although the Lyapunov exponents of a chaotic system are the same for nearly all initial conditions, the

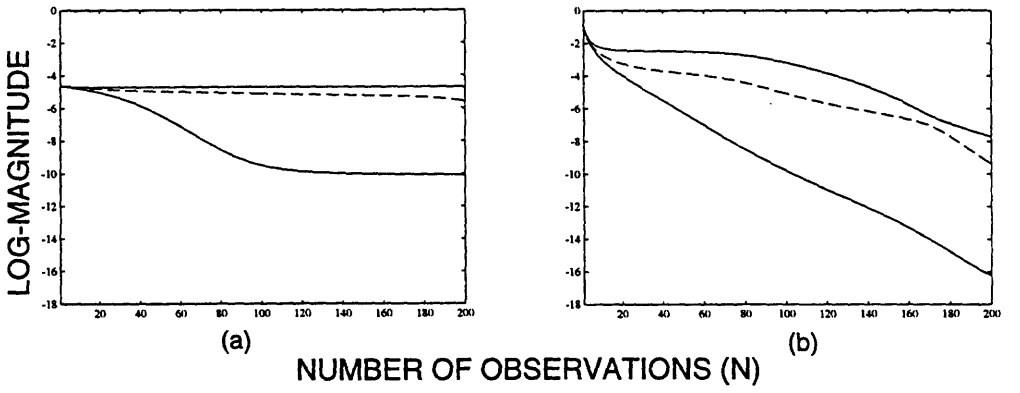

Figure 5-18: Eigenvalues of random Cramer-Rao bound for initial condition estimators of time-sampled Lorenz flow with observation set $Y(0, N)$ and input SNR of 20 dB. (a) $\gamma = .1$; (b) $\gamma = 1$.

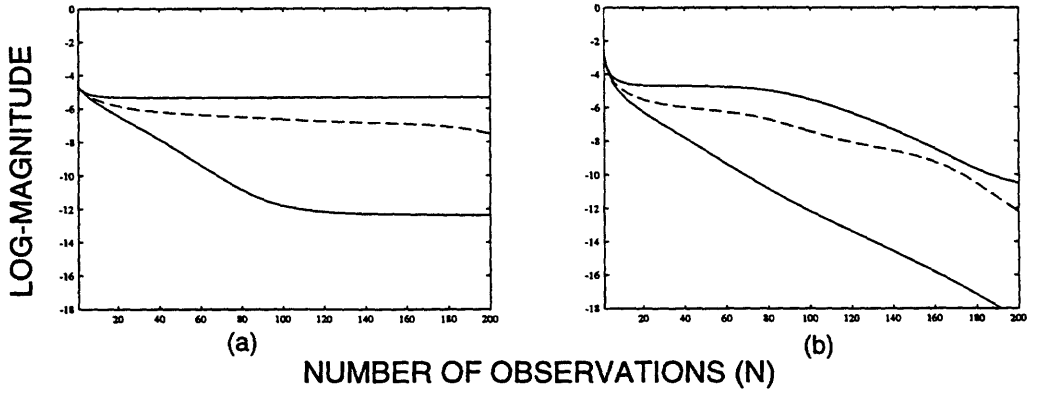

Figure 5-19: Eigenvalues of random Cramer-Rao bound for initial condition estimators of time-sampled Lorenz flow with observation set $Y(0, N)$ and input SNR of 30 dB. (a) $\gamma = .1$; (b) $\gamma = 1$.

local manifolds or directions associated with these exponents differ at each point. With observations of future state values and a fixed initial condition $\boldsymbol{x}(0)$, the Fisher information (as given by $\boldsymbol{J}(\boldsymbol{x}(0))$) along the directions associated with negative Lyapunov exponents remains small as the number of future observations increases, with a similar relation among the directions associated with positive Lyapunov exponents and past observations. Equivalently, $\boldsymbol{J}(\boldsymbol{x}(0))$ has bounded eigenvalues, one for each negative Lyapunov exponent, which thus have bounded, nonzero inverses, even as the number of future observations becomes large. However, the eigenvectors associated with these eigenvalues differ for different values of $\boldsymbol{x}(0)$. All other eigenvalues increase without bound as the number of future observations increases. When the expectation is taken over $\boldsymbol{x}(0)$ and there is thus an averaging of the matrices $\boldsymbol{J}(\boldsymbol{x}(0))$, the resulting averaged Fisher information is not small in any direction for the larger value of γ. Equivalently, both the large and small eigenvalues of the individual matrices $\boldsymbol{J}(\boldsymbol{x}(0))$ influence each eigenvalue of $E_{\boldsymbol{x}(0)}\{\boldsymbol{J}(\boldsymbol{x}(0))\}$ with computer experiments suggesting that each of these eigenvalues increases as the number of future observations increases. Since the inverses of these eigenvalues (incremented by γ^{-2}) are the eigenvalues of $\boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0))$, the result is that all eigenvalues of $\boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0))$ decrease with increasing N. In contrast, for the smaller value of γ, the local manifolds associated with Lyapunov exponents for the values of $\boldsymbol{x}(0)$ used in the averaging differ very little, so that the average of the matrices $\boldsymbol{J}(\boldsymbol{x}(0))$ has similar eigenvalue properties as each of the matrices $\boldsymbol{J}(\boldsymbol{x}(0))$ individually.

This averaging of information matrices can lead to undesirable behavior by the Cramer-Rao bound. Figures 5-20 (a) and (b) depict the superimposed eigenvalue sums of $\boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0))$ for the two values of γ used for the earlier figures and with input SNRs of 20 dB and 30 dB, respectively. In both figures, there is a threshold value of N above which the eigenvalue sum is smaller for the larger value of γ. In addition, this threshold is smaller with the larger input SNR value. The figures indicate that when N increases beyond a threshold, the bound on the MSE is smaller when there is greater *a priori* uncertainty in the initial condition $\boldsymbol{x}(0)$. If in fact the bound were achievable, the nonsensical implication would be that achievable state estimator performance increases with decreasing *a priori* knowledge of the actual value of the initial condition. Furthermore, this undesirable behavior of the bound becomes more pronounced as the input SNR increases. Consequently, the random Cramer-Rao bound on the performance of initial condition estimators of dissipative, chaotic

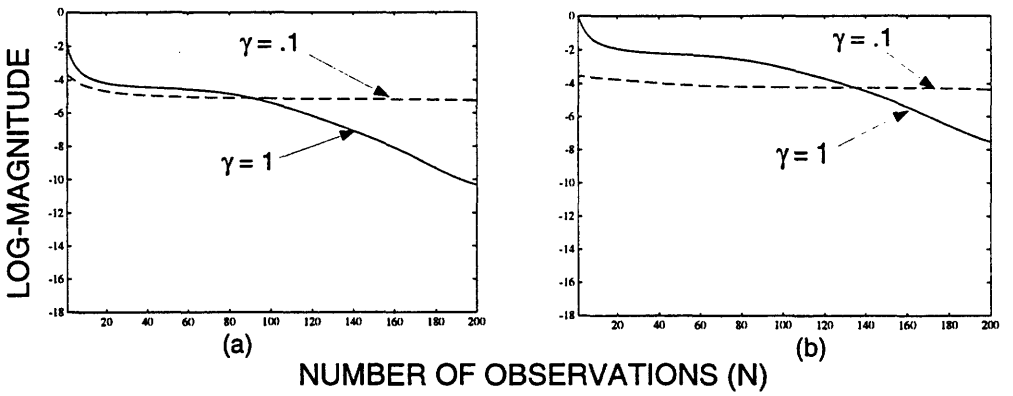

Figure 5-20: Summed eigenvalues of random Cramer-Rao bound for initial condition estimators of time-sampled Lorenz flow with observation set $Y(0, N)$. (a) Input SNR = 30 dB; (b) Input SNR = 20 dB.

maps is least effective with input SNRs for which the Cramer-Rao bound is generally most effective.

5.4.2 Weiss-Weinstein Bound

Just as the nonrandom Cramer-Rao bound has a random counterpart, the Barankin bound has random counterparts as well [9, 88, 89, 90], with the most general of these counterparts being the Weiss-Weinstein class of performance bounds for estimators of random parameters. In contrast to the probabilistic setting for the Barankin bound provided in Section 5.2.2, the probabilistic setting for the Weiss-Weinstein involves two probability spaces—one the observation space (Ω, β, μ) (as with the Barankin bound), and the other the parameter space (Θ, η, ν), where $\Theta = \{\theta\}$ is a set of parameter values, η is a σ-algebra of subsets of Θ, and ν is a probability measure defined on η. An underlying assumption is that the joint probability measure on the two spaces is absolutely continuous with respect to the product measure, so that the joint density function $p(\omega, \theta)$ exists. As shown in [90] (with slightly different notation), for any real-valued function $\hat{g}(\omega)$ that is measurable on Ω the following inequality holds for all finite n, real constants a_i, offsets z_i, and exponents s_i satisfying $0 \leq s_i \leq 1$, for $i = 1, \cdots, n$:

$$E_{\omega,\theta}\left\{(\hat{g}(\omega) - g(\theta))^2\right\} \geq \frac{\left[\sum_{i=1}^{n} a_i \, E_{\omega,\theta}\left\{[g(\theta - z_i) - g(\theta)] L^{1-s_i}(\omega, \theta - z_i, \theta)\right\}\right]^2}{E_{\omega,\theta}\left\{\left\{\sum_{i=1}^{n} a_i \, [L^{s_i}(\omega, \theta + z_i, \theta) - L^{1-s_i}(\omega, \theta - z_i, \theta)]\right\}^2\right\}} \tag{5.42}$$

where

$$L(\omega, \theta_1, \theta_2) \equiv \frac{p(\omega, \theta_1)}{p(\omega, \theta_2)}. \tag{5.43}$$

for $\theta_1, \theta_2 \in \Theta$ and where $E_{\omega,\theta}$ denotes expectation over the joint density $p(\omega, \theta)$. The above inequality, the Weiss-Weinstein bound, is the counterpart to (5.19), the inequality corresponding to the Barankin bound. Note that whereas constant test points ω_i corresponding to other parameter values are used in the Barankin bound, constant offsets z_i are used in the Weiss-Weinstein bound.

Appendix A provides the equations for a restricted form of the Weiss-Weinstein bound for vector-valued parameters for the problem scenario of interest in this chapter. As shown in the appendix, analogous to the restricted form of the Barankin bound considered earlier in the chapter, this restricted form of the Weiss-Weinstein bound is expressible as a sum of two components—one the inverse of the random Fisher information matrix, and the other a matrix which depends on the test offsets, the observation noise covariance matrix, and the number of observations. Figure 5-21 depicts the random Cramer-Rao and restricted Weiss-Weinstein bounds for the same scenario used for Figure 5-20.

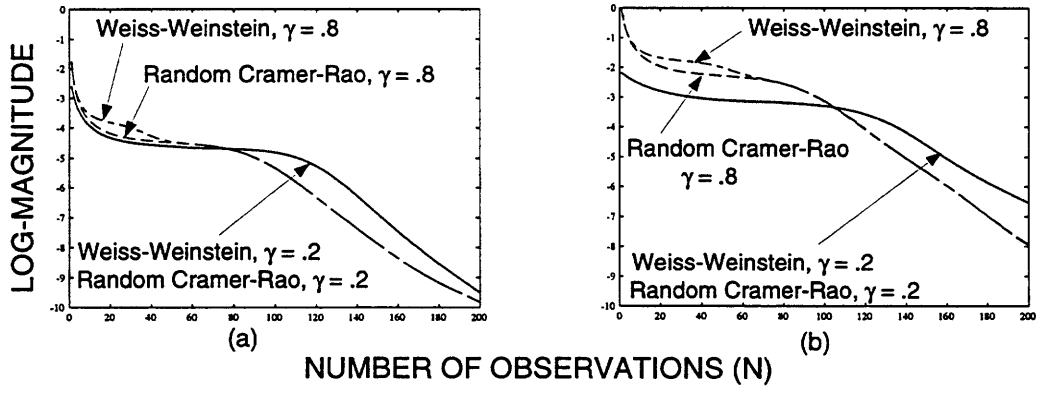

Figure 5-21: Summed eigenvalues of restricted Weiss-Weinstein and random Cramer-Rao bound for initial condition estimators of time-sampled Lorenz flow with observation set $Y(0, N)$. (a) Input SNR = 30 dB; (b) Input SNR = 20 dB.

A single test offset was used in the Weiss-Weinstein bound, with the offset carefully chosen by trial-and-error to yield as tight a bound as possible. Additional experiments with more than one test offset offered little if any improvement. As indicated by the figures, the undesirable behavior exhibited by the random Cramer-Rao bound is also exhibited by the restricted Weiss-Weinstein bound, in the sense that there is a threshold value of N above which the bound on the MSE is smaller when there is greater *a priori* uncertainty in the

initial condition $\boldsymbol{x}(0)$. As such, the undesirable behavior of the random Cramer-Rao bound with larger input SNRs is not avoided with generalizations of the bound.

The question arises as to the existence of a performance bound for estimators of random parameters which avoids the undesirable behavior of the random Cramer-Rao bound. It is straightforward to derive a rate-distortion bound for the problem of interest here either by specializing the more general bounds introduced in [28, 92] or by using the theory developed in [7] as a foundation. Experiments have suggested that the resulting bound avoids the undesirable behavior of the random Cramer-Rao bound but that the bound is a much weaker bound than the Cramer-Rao bound except with extremely small input SNRs and thus has little value with the input SNRs of interest here. Alternatively, the random Cramer-Rao bound does not need to be incorporated in the Weiss-Weinstein bound as was done for the results presented above. However, computer experiments have indicated that with input SNRs above 20 dB, the Weiss-Weinstein bound is an extremely weak bound when it does not incorporate the random Cramer-Rao bound.

Chapter 6

MC Maps and Signal Synthesis

6.1 Introduction

Thus far in this thesis, we have focused on dissipative, chaotic systems and state estimation with these systems. Various topological and ergodic properties of dissipative, chaotic systems have been considered, including the presence of both positive and negative Lyapunov exponents, boundedness of attractors, and the existence of invariant measures; and the influence of these properties on achievable state estimator performance has been addressed. Until now, we have not considered properties unique to individual chaotic systems, nor have we discussed how one might synthesize a chaotic signal with specified properties.

In contrast to the previous chapters, this chapter focuses on signal synthesis, in particular the synthesis of maps with properties that facilitate the detection and estimation of noise-corrupted orbit segments generated by these maps. We introduce a class of maps that are amenable to analysis and although deterministic, generate potentially useful random processes. In the next chapter, we exploit this random-process-generation property of these maps to derive computationally efficient, practical, optimal and suboptimal detection and estimation algorithms involving these maps, and we briefly speculate on the use of these maps and detection algorithms for secure communication.

Unlike the maps considered in Chapters 4 and 5, the maps considered in this chapter and the next are neither dissipative nor invertible, and the attractors are not fractals but simple subsets of $\mathcal{R}$. We also seldom use the word chaos in this chapter, thereby avoiding the difficulties encountered in Chapter 2 in properly defining the word and in determining whether a given system satisfies a given definition. However, many of the maps considered in

this chapter have a positive Lyapunov exponent, bounded orbits, and sensitive dependence on initial conditions and would thus satisfy many definitions of chaos.

Our interest in the maps discussed in this chapter arose in an effort to better understand the behavior of certain state estimation algorithms involving Kalman filtering and hidden Markov modeling with dissipative, chaotic maps. However, because of the rich set of properties these simple, one-dimensional maps exhibit, the maps themselves have potential, practical value and are worthwhile to investigate independently of dissipative systems.

The chapter begins by introducing MC maps, a class of piecewise linear maps of the unit interval onto itself which give rise to Markov chains and continues by presenting previously reported and newly discovered properties of MC maps. Finally, the chapter discusses the use of MC maps for synthesizing maps having specified stationary PDFs and for synthesizing multidimensional maps which also give rise to Markov chains.

A notational convention used in the chapter and the next is that for real numbers x and y satisfying $x < y$, $[x, y]$ and $]x, y[$ respectively denote the closed and open intervals of the real line R with x and y as endpoints. Similarly, $]x, y]$ and $[x, y[$ respectively denote left-open, right-closed and left-closed, right-open subintervals of R. Also, λ denotes Lebesgue measure, whereas μ denotes an arbitrary measure.

6.2 Markov Maps and MC Maps

The maps of principal interest in this chapter and the next comprise a subset of a much larger class of maps of intervals of the real line to themselves known as Markov maps. The formal definition of a Markov map is the following:

Markov Map: [12, 25] A piecewise continuous map f of an interval $I = [i_0, i_n]$ to itself for which there exists a set of points $P \equiv \{i_1, i_2, \cdots, i_{n-1}\}$ known as *partition points* satisfying $i_0 < i_1 < \cdots < i_{n-1} < i_n$ and such that the following two conditions hold:

1. For $j = 0, 1, \cdots, n-1$, the restriction of f to the open subinterval $]i_j, i_{j+1}[$ is a homeomorphism (i.e., a continuous, invertible mapping with continuous inverse) onto another subinterval $]i_{k(j)}, i_{l(j)}[$, where $i_{k(j)}$ and $i_{l(j)}$ are elements of the set $\{i_0, i_1, \cdots, i_n\}$ and $i_{k(j)} < i_{l(j)}$.

2. $f(P) \subset P$, which means that partition points are mapped to partition points.

3. $f(i_j^-), f(i_j^+) \in P$, where $f(i_j^-) \equiv \lim_{\delta \to 0} f(i_j - \delta)$, $f(i_j^+) \equiv \lim_{\delta \to 0} f(i_j + \delta)$, and $\delta > 0$. This means that the left and right limits of f evaluated at each partition point are also partition points.

Intuitively, this abstract definition means that a Markov map is a piecewise continuous map of an interval I to itself for which one can find a partition of I into (nonoverlapping) subintervals such that each subinterval is mapped *nicely* onto a union of other subintervals in the partition, and such that the endpoints of these subintervals are mapped onto endpoints of other subintervals in the partition. Implicit in the above definition is the fact that f gives rise to the deterministic state equation $x(n+1) = f(x(n))$ where $x(n) \in I$. We provide examples of Markov maps later in the section.

A *Markov partition* is any finite partition of the interval $I = [i_0, i_n]$ into subintervals for which f satisfies the three conditions for a Markov map. Since partition elements are subintervals of R, we use the terms *partition elements* and *subintervals* interchangeably in this chapter to refer to these elements. A Markov map may have many Markov partitions, and in general it is difficult both to determine if a given map is a Markov map and to find a Markov partition for a given Markov map. An important consideration, which underlies the detection algorithms introduced later in the chapter, is that certain Markov maps have an infinite set of Markov partitions that are straightforward to determine.

Of particular interest in this chapter is a small subset of Markov maps for which each element f of the subset satisfies the following, additional constraint:

- f is piecewise linear and there exists a Markov partition for which f is affine on each partition element. A piecewise linear map of an interval I to itself is a map that is affine on each subinterval in a set of subintervals partitioning I. An affine map is a map of the form $f(x) = \tau x + \beta$, where τ and β are real-valued constants.

We use the term *MC map* to denote an element of this subset of Markov maps, because as shown later in the section, such a map gives rise to homogeneous, finite-state Markov chains. Also, we use the term *EMC map* to denote an MC map which satisfies the following additional constraint

- The map is *eventually locally expanding* in the sense that there exists an integer n such that $|D\{f^n(x)\}| > 1$ at all differentiable points of f^n.

Because it satisfies this additional constraint, an EMC map has a positive Lyapunov exponent and in general exhibits sensitive dependence on initial conditions. A fact we exploit later in the chapter is that this additional constraint is satisfied by any piecewise linear map for which the slope of each affine segment is an integer with absolute value greater than one.

Figures 6-1 (a) and (b) depict two EMC maps of the unit interval. Since $x(n+1) = f(x(n))$, the figures depict the function f defining the map. Figure 6-2 (a) depicts an orbit segment $\{x(i)\}_{i=0}^{N}$ generated by the map in Figure 6-1 (a), and Figure 6-2 (b) depicts an orbit segment generated by the map in Figure 6-1 (b).

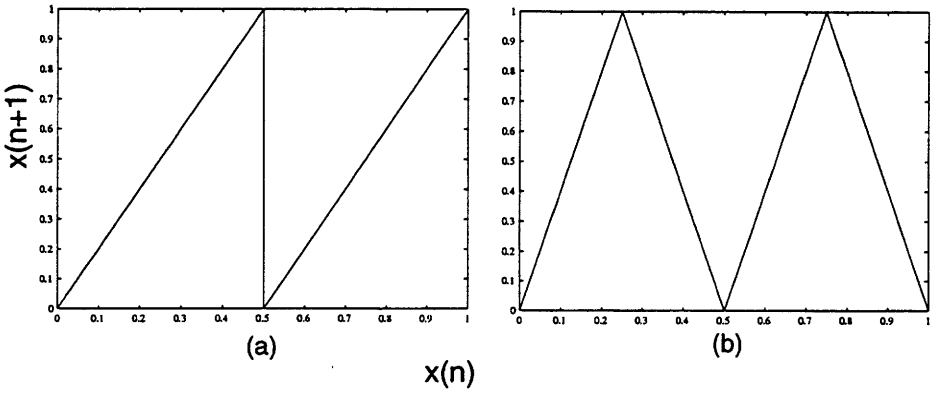

Figure 6-1: Two EMC maps

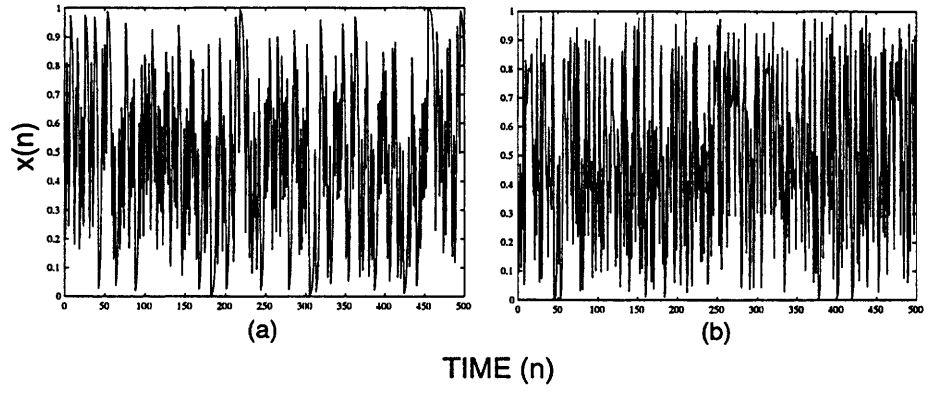

Figure 6-2: Typical orbit segments for respective maps in Figures 6-1 (a) and (b)

A Markov partition for the map shown in Figure 6-1 (a) is given by any division of the unit interval into $2N$ equal length subintervals, where N is a positive integer. Similarly, a Markov partition for the map shown in Figure 6-1 (a) is given by any division of the unit interval into $4N$ equal length subintervals. For example, the partition given by the four

equal length subintervals $\{[0,.25[,[.25,.5[,[.5,.75[,[.75,1]\}$ is a Markov partition for each map.

A useful property of MC maps, which we exploit later in this chapter, is that under certain conditions on the distribution of the initial condition $x(0)$, the maps give rise to Markov chains. As shown in [12, 61], a Markov map f of the unit interval gives rise to a Markov chain if the initial condition has a constant PDF over the unit interval and there exists a Markov partition $P = \{I_j\}$ satisfying the following two constraints:

1. $f(x) = \tau_{jk}\, x + \beta_{jk}$, for $x \in I_{jk}$ and real constants τ_{jk}, β_{jk}
2. $f(I_{jk}) = I_k$ almost everywhere if $I_{jk} \neq \emptyset$

where I_{jk} denotes the points of I_j which are mapped to I_k by f. By definition, an MC map satisfies these two constraints and thus gives rise to a Markov chain if the initial condition is appropriately chosen. As we show later, the constraint on the PDF of the initial condition can be relaxed. We also show that many other Markov maps both one-dimensional and multidimensional, which do not satisfy the above constraints, give rise to Markov chains as well.

For a given MC map defined on the unit interval $[0,1]$ and a given Markov partition $\{I_j\}$ for which the above two constraints are satisfied, each partition element I_j corresponds to a state S_j of a Markov chain. Since the restriction of f to I_j is an affine transformation, then I_{jk} is either a subinterval of I_j or the empty set. In fact, a readily verified relation we use later in the chapter is the following: $I_{jk} = I_j \cap f^{-1}(I_k)$. The transition probability from state S_j to state S_k equals the fraction of points in I_j that are mapped to I_k, which is given by $\lambda(I_{jk})/\lambda(I_j)$ where λ is Lebesque measure.

For the Markov maps shown in Figure 6-1 (a) and (b), with the Markov partition given by $\{[0,.25[,[.25,.5[,[.5,.75[,[.75,1]\}$, the matrices of state transition probabilities, hereafter referred to as transition probability matrices (TPMs), are the following, respectively:

$$\begin{bmatrix} .5 & .5 & 0 & 0 \\ 0 & 0 & .5 & .5 \\ .5 & .5 & 0 & 0 \\ 0 & 0 & .5 & .5 \end{bmatrix} \qquad \begin{bmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{bmatrix}. \tag{6.1}$$

The dynamics of the Markov chain arise as follows. Let $\{I_j\}$ denote the elements of a

Markov partition for f, and let $\{S_j\}$ denote the states of the Markov chain corresponding to this partition, where state S_j is associated with partition element I_j. Also, let $x(n)$ denote the state or orbit point of f at time n, i.e., $x(n) = f^n(x(0))$. To avoid confusion or ambiguity, we henceforth use the word state only in reference to Markov chains. If $x(n) \in I_j$, the Markov chain is said to be in state S_j at time n. If the initial condition is a random variable with appropriate distribution (e.g., constant PDF), the state sequence that arises as a result of this mapping between orbit points and states is a first-order Markov process with transition probabilities defined as above.

More formally, consider the random sequence $\{S(n)\}$ where $S(n)$ denotes the state corresponding to the partition element in which $x(n)$ lies (i.e., $S(n) = S_j$ if $x(n) \in I_j$). Also, let $P(S(0))$ denote the initial state distribution given by $P(S(0) = j) = \lambda(I_j)$, and let $P(S(0), S(1), \cdots, S(N))$ denote the joint probability of $S(0), S(1), \cdots, S(N)$. The condition on $P(S(0))$ implies that the initial condition $x(0)$ is a random variable with constant PDF over the unit interval. Under these conditions, the following Markov property holds:

$$P(S(0), S(1), \cdots, S(N)) = P(S(0))\, P(S(1)|S(0)) \cdots P(S(N)|S(N-1)) \tag{6.2}$$

where

$$P([S(i) = k]|[S(i-1) = j]) = \frac{\lambda(I_{jk})}{\lambda(I_j)} \tag{6.3}$$

In the next chapter, we exploit the relation between MC maps and Markov chains to derive algorithms for detecting noise-corrupted orbit segments generated by these maps.

EMC maps are among the few maps exhibiting the properties associated with chaos that are also amenable to analysis. Analysis of these maps is facilitated by their piecewise linearity. Nonetheless, these maps have many interesting, potentially useful properties and may be useful *building blocks* for other signals. In the next 4 sections, we consider properties of EMC maps and more generally MC maps, which are relevant to detection and estimation applications involving these maps. Among the properties we consider is the relative ease in synthesizing an MC map with integer-valued slopes for the affine segments, which gives rise to any specified Markov chain having a specified TPM with rational-valued entries. The significance of integer-valued slopes is that any MC map having this property gives rise to arbitrarily fine Markov partitions with equally sized partition elements, a fact we prove later in the chapter. We also merge previously published results to establish a close relation

between the properties of EMC maps and the Markov chains they give rise to. In particular, we show that ergodicity of an EMC map is equivalent to irreducibility of the TPMs of all Markov chains the map gives rise to and exactness of a EMC map is equivalent to primitivity of the TPMs of all Markov chains the map gives rise to. We also show that the Markov chain property of EMC maps is preserved under homeomorphisms, a result which allows the synthesis of a Markov map which gives rise to any specified Markov chain and which has any specified stationary PDF. Finally, we show how one can use this homeomorphism property to design nontrivial multidimensional maps which give rise to Markov chains.

In the next section, we focus on the relation between MC maps and the Markov chains they give rise to, whereas in Section 6.4 we deal with the relation between these maps and their associated Markov partitions. Section 6.5 is devoted to the ergodic properties of EMC maps, and Section 6.6 deals with the synthesis of multidimensional MC maps. Various theorems and propositions are presented, with proofs provided for those that are new. Several of the proofs are informal, with the essential elements of the formal proofs provided but the excessive, unrevealing details omitted.

6.3 EMC Maps and Markov Chains

As originally shown in [43, p.294:Theorem 5] and later in [46, 80], given the transition probability matrix (TPM) of any homogeneous, finite-state Markov chain and any vector of nonzero, initial state probabilities, one can synthesize a piecewise linear map of the unit interval onto itself which gives rise to that Markov chain. The map will be a Markov map, but may or may not be an MC map depending upon the specified Markov chain. Thus, Markov maps are generators of homogeneous, finite-state Markov chains. We briefly outline the design process, primarily to clarify the relation between piecewise linear Markov maps and Markov chains. To facilitate an understanding of the design process, we apply the method to a specific example while outlining the process.

Let m denote the number of states of the desired Markov chain, and let the row vector of state probabilities at time n be denoted $\boldsymbol{\Pi}(n) = [\pi_1(n), \pi_2(n) \cdots, \pi_m(n)]$, where $\pi_i(n)$ denotes the probability of being in state i at time n. With this notation, $\boldsymbol{\Pi}(0)$ denotes the vector of initial state probabilities. Also, let the $m \times m$ TPM be denoted $\boldsymbol{P} = [p_{ij}]_{i,j=1}^{m}$ where p_{ij} denotes the probability of transitioning from state i to state j in one time step.

For our example, we let $m = 3$, $\boldsymbol{\Pi}(0) = [.25, .25, .5]$, and we let $\boldsymbol{P}$ be given by

$$\begin{bmatrix} .5 & .5 & 0 \\ 0 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}. \tag{6.4}$$

Now perform the following sequence of steps:

1. Partition the unit interval into m consecutive subintervals $\{I_j\}_{j=1}^m$, where I_j has length $\pi_j(0)$. Let $e_{j,l}$ and $e_{j,r}$ respectively denote the left and right endpoints of I_j. (Note that $e_{j,r} = e_{j+1,l}$.) For our example, the set of (closed) subintervals $[e_{j,l}, e_{j,r}]$ is given by $\{[0, .25], [.25, .5], [.5, 1]\}$.

2. Partition each subinterval I_j into m consecutive subintervals $\{I_{jk}\}_{k=1}^m$ where I_{jk} has length $\pi_j(0)\, p_{jk} = \lambda(I_j)\, p_{jk}$, which is the product of the transition probability from state j to state k and the length of subinterval I_j. Also let $e_{jk,l}$ and $e_{jk,r}$ respectively denote the left and right endpoints of I_{jk}. (Note that if $p_{jk} = 0$, the subinterval I_{jk} has zero length. Zero length subintervals are included here only as a notational aid to simplify the use of indices). For our example, the sets of subintervals $[e_{jk,l}, e_{jk,r}]$ for fixed j and increasing k are given by

$$\{[e_{1\cdot,l}, e_{1\cdot,r}]\} = \{[0, .125], [.125, .25], [.25, .25]\} \tag{6.5}$$

$$\{[e_{2\cdot,l}, e_{2\cdot,r}]\} = \{[.25, .25], [.25, .25], [.25, .5]\} \tag{6.6}$$

$$\{[e_{3\cdot,l}, e_{3\cdot,r}]\} = \{[.5, \frac{2}{3}], [\frac{2}{3}, \frac{5}{6}], [\frac{5}{6}, 1]\}. \tag{6.7}$$

3. Now consider a map f from the unit interval to itself as a function in the (x, y)-plane which maps each point $x_0 \in [0, 1]$ to some point $y_0 \in [0, 1]$. Let (x_0, y_0) denote the relation $y_0 = f(x_0)$. For each subinterval I_{jk} with nonzero length, draw the line segment in the (x, y)-plane with left endpoint given by the (x, y) pair $(e_{jk,l}, e_{k,l})$ and right endpoint given by the (x, y) pair $(e_{jk,r}, e_{k,r})$. This corresponds to a linear mapping of the subinterval I_{jk} (on the x-axis) onto the subinterval I_k (on the y-axis). Figure 6-3 depicts the resulting piecewise linear map for our example. Special care must be taken at each discontinuity point to ensure that the point maps to a single point.

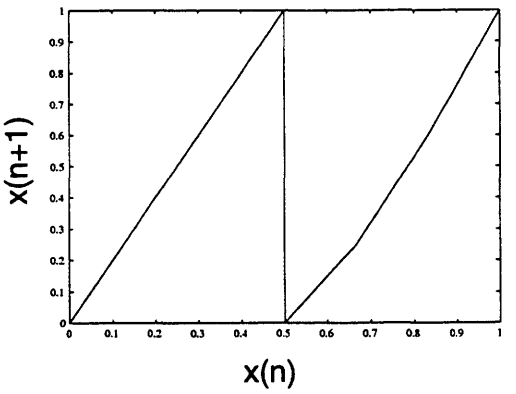

Figure 6-3: Synthesized MC map.

The critical aspects of the synthesis procedure are in specifying the lengths of the various subintervals and in ensuring that each subinterval I_{jk} maps linearly onto I_k. Note that since I_{jk} has length $\pi_j(0)\,p_{jk} = \lambda(I_j)\,p_{jk}$ and I_k has length $\pi_k(0) = \lambda(I_k)$, it follows that $f'(I_{jk})$, the slope of the affine mapping of I_{jk} onto I_k, is given by

$$f'(I_{jk}) = \frac{\pi_k(0)}{\pi_j(0)\,p_{jk}} = \frac{\lambda(I_k)}{\lambda(I_j)\,p_{jk}} \tag{6.8}$$

The synthesis procedure is not the only one possible. For example, the substitution of the line segment with left-right endpoint pairs $(e_{jk,l}, e_{k,r})$ and $(e_{jk,r}, e_{k,l})$ for the segment with left-right endpoint pairs $(e_{jk,l}, e_{k,l})$ and $(e_{jk,r}, e_{k,r})$ yields the same results.

The subintervals $\{I_j\}$ are the m states of the Markov chain. The state sequence arising with almost all initial conditions $x(0)$ for the induced dynamical system $x(n+1) = f(x(n))$ is a sample path of the desired Markov chain, provided that $x(0)$ is a random variable with constant PDF over the unit interval. As discussed earlier in this thesis, randomness in a deterministic, chaotic system is due solely to randomness in the initial condition. However, this randomness constraint on the initial condition is necessary only to ensure the initial state probabilities have the desired values. Consider the situation in which we only observe the partition element in which $x(n)$ lies at each time n, or equivalently we know the state sequence of the Markov chain and nothing else. Then, regardless of the distribution of the initial condition, the transitions between states will be random events with the desired transition probabilities. In other words, with restricted knowledge about the orbit point $x(n)$ at each time n, the original, deterministic process becomes a random process with the desired Markov chain structure.

Although the map f that results from the synthesis procedure is an MC map, the partition given by the set of intervals $\{I_j\}$ is not a Markov partition for the map since the restriction of f to each of these intervals is not an affine mapping, but instead a piecewise linear mapping. However, the finer partition given by the set of subintervals $\{I_{jk}\}$ is in fact a Markov partition for f.

We now extend the relation between Markov chains and MC maps in two ways, with both extensions useful for the detection and estimation applications considered in the next chapter. First, we have the following result that simplifies the selection of Markov partitions.

Proposition 1: Given a TPM with rational-valued elements and a vector of rational-valued, nonzero initial state probabilities for a Markov chain, one can synthesize an MC map and if desired an EMC map which gives rise to this Markov chain, with each of the affine segments of the map having *integer-valued* slope.

Proof: (see Appendix B)

Second, the procedure discussed earlier for synthesizing Markov maps which give rise to Markov chains requires that the initial condition $x(0)$ have a constant PDF over the unit interval, that the initial state probabilities all be nonzero, and that the length of each interval I_j be equal to the initial probability of the corresponding state. Each of these requirements can be relaxed and the Markov chain property still hold, as indicated by the following corollary.

Corollary 1: The Markov chain property of the Markov map designed with the procedure discussed earlier still holds if the initial condition $x(0)$ is a random variable with constant PDF over each interval I_j, but the constant value need not be the same for different intervals. In this case, the initial state probability for S_j, the state corresponding to interval I_j, is given by the product of the interval length $\lambda(I_j)$ and the constant value associated with that interval by the PDF of the initial condition.

Proof: We omit a formal proof. Instead, we note that the proof of Proposition 1 still holds with only minor changes when the PDF of the initial condition is constant over each subinterval I_j. The state transition probabilities remain the same since they are determined by ratios of subinterval lengths, i.e., $p_{jk} = \frac{\lambda(I_{jk})}{\lambda(I_j)}$. (Note that this corollary may in part be implied by Corollary 1.2 in [61]).

In light of the above corollary, we are free to arbitrarily specify the lengths of the

intervals corresponding to each state of the Markov chain provided that each interval has nonzero length and the distribution of $x(0)$ is chosen appropriately. This result is useful if we wish to cascade two MC maps in the sense that we generate an orbit segment $\{x(i)\}_{i=0}^{N}$ with one map and then use the final orbit point $x(N)$ as the initial condition for the second map. As we discuss Section 6.5, a subclass of EMC maps have unique, stationary PDFs that are constant over the partition elements of suitably chosen Markov partitions. As a result, if the two EMC maps are synthesized appropriately, the state sequence arising from the second map with initial condition $x(N)$ is a sample path of a Markov chain as well.

6.4 Markov Partitions

The questions arise as to how one goes about choosing Markov partitions for MC maps and how one determines if a piecewise linear map is a Markov map or an MC map. The selection of Markov partitions for MC maps is an important consideration when these maps are used for detection and estimation applications.

Unfortunately, there are no comprehensive answers to these questions but instead a host of partial answers. For example, one necessary condition for a piecewise linear map to be a Markov map is that the endpoints of the affine segments be periodic or eventually periodic points of the map. The necessity of this condition is a consequence of the fact that a Markov map must have a Markov partition for which the restriction of the map to each partition element is an affine transformation. Thus, the endpoints of the affine segments must be partition points for such a partition. Since partition points are required to map to partition points and each partition is required to have a finite number of elements, it follows that these endpoints must be periodic or eventually periodic points of the map. Otherwise, there would be an infinite number of partition points and concomitantly an infinite number of partition elements.

One special case in which the periodic points are particularly simple to find is when the slope of each affine segment is integer-valued. As shown in [78], in this special case, all rational points are periodic or eventually periodic. If in addition the endpoints of the domains of the affine segments and the images of these endpoints are rational-valued, the following, stronger result holds as well.

Proposition 2: For any piecewise linear map f of the unit interval to itself for which

the slope of each affine segment is integer-valued and the endpoints of the affine segment domains along with the images of these endpoints are rational-valued, there exists a uniform Markov partition of the map. Furthermore, any uniform refinement of such a partition is also a Markov partition. By a *uniform Markov partition*, we mean a Markov partition for which each partition element has the same length. By a *uniform refinement* of a uniform Markov partition, we mean any uniform Markov partition for which the set of partition points include those of the original uniform partition.

Proof: (See Appendix B)

Earlier we showed that one can synthesize an MC map with integer-valued slopes that gives rise to any finite-state, homogeneous Markov chain with rational-valued transition probabilities. As indicated by the above proposition, for these maps one can always find a sequence of increasingly finer, uniform Markov partitions.

When given both an MC map f for which the slopes of the affine segments are not all integer-valued and a Markov partition for the map, the question arises as to how one goes about finding Markov partitions that are refinements of this partition. A simple approach is to use successive inverse images of the original partition points as additional partition points. These additional points get mapped by f or compositions of f to the original partition points, and thus satisfy the requirement that partition points be mapped to partition points. This approach was used in [14, 46] to approximate arbitrary, monotonically expanding, one-dimensional maps by piecewise linear Markov maps.

6.5 Ergodic Properties of EMC Maps

As one might expect, there is a strong relation between the ergodic properties of MC maps and the Markov chains which they give rise to. In this section, we explore this relation and in particular show that ergodicity of an EMC map is equivalent to irreducibility of the TPMs of the Markov chains the map gives rise to, and exactness of a EMC map is equivalent to primitivity of the TPMs of the Markov chains the map gives rise to. As we discuss later in the section, the practical relevance of stationary PDFs in the context of EMC maps is that they facilitate the synthesis of maps, not necessarily EMC maps, which give rise to any specified Markov chain and and any stationary PDF. We do not consider the spectral properties of EMC maps, a topic closely related to that of stationary PDFs.

The reader is referred to [37] for an exposition on this topic.

6.5.1 Ergodic Theory Fundamentals

Before discussing the ergodic properties of EMC maps, we briefly review several relevant concepts from ergodic theory not considered in Chapter 2. Much of the information is extracted from [50, 55, 87] to which the reader is referred for additional details.

As in Section 2.3.2, we consider a measure space (X, β, μ) where X is a set, β is a σ-algebra of subsets of X, and μ is a measure on β. Of special interest is when X is the unit interval I (or an arbitrary interval on the real line) and β is the Borel σ-algebra on I, which is defined as the smallest σ-algebra that includes the open intervals on the real line intersected with I. Also, of special interest is when μ is either Lebesgue measure, or a measure that is absolutely continuous with respect to Lebesgue measure, which as a consequence has a corresponding probability density function (PDF). In other words, we are principally interested in very simple measure spaces with probability measures having corresponding PDFs.

As noted in our earlier discussion of ergodic properties, given a measure space (X, β, μ), a transformation $f : X \to X$ is *measurable* if $f^{-1}(B) \in \beta$ for every $B \in \beta$. This means that the inverse image of every element of β is also in β. A measurable transformation is *nonsingular* if $\mu(f^{-1}(B)) = 0$ for every $B \in \beta$ for which $\mu(B) = 0$. A measurable transformation is *measure-preserving* if $\mu(f^{-1}(B)) = \mu(B)$; μ is then said to be an *invariant measure* of f. Our principal interest is in measure-preserving transformations for which the invariant measure has a corresponding PDF.

Given a measure space (X, β, μ), let $L^1(\mu)$ denote the set of all absolutely integrable functions on (X, β, μ), where an absolutely integrable function g on (X, β, μ) is a real-valued function (i.e., $g : X \to \mathcal{R}$) that is measurable on (X, β, μ) and satisfies

$$\int |g(x)| \, d\mu(x) < \infty. \tag{6.9}$$

Each $h \in L^1(\mu)$ which satisfies $h(x) \geq 0$, $\forall x \in X$ and $\int h(x) \, d\mu(x) = 1$ is known as a *density function* on (X, β, μ). When μ is Lebesgue measure λ, such a function is simply a PDF familiar to non-mathematicians. Note that each density function h induces a probability measure μ_h defined by $\mu_h(B) = \int_B h(x) \, d\mu(x)$ for each $B \in \beta$.

For each nonsingular transformation $f : X \to X$ on (X, β, μ), there is a unique operator $P_f : L^1(\mu) \to L^1(\mu)$ known as the *Frobenius-Perron operator* which satisfies

$$\int_B P_f(p(x))\, d\mu(x) = \int_{f^{-1}(B)} p(x)\, d\mu(x) \tag{6.10}$$

for each $B \in \beta$ and $p \in L^1(\mu)$ [50, 55]. This rather abstract definition of the Frobenius-Perron operator has a simple, intuitive interpretation when X is the real line $\mathcal{R}$, β is the Borel σ-algebra, μ is Lebesgue measure, and $p(x)$ is a PDF. With these restrictions, $P_f(p(x))$ is the PDF induced by the transformation f. In other words, $P_f(p(x))$ is the PDF which when integrated over the set $f(B)$ for each $B \in \beta$, gives the same value as is obtained by integrating $p(x)$ over B. If f is invertible and differentiable and we let $y = f(x)$, the defining expression for $P_f(p(x))$ reduces to the following expression found in many probability textbooks (e.g., [68, p. 118]) for the PDF $p_Y(y)$ induced by f:

$$p_Y(y) = \frac{p(f^{-1}(y))}{|f'(f^{-1}(y))|}. \tag{6.11}$$

where $f'(f^{-1}(y))$ denotes the derivative of f evaluated at $f^{-1}(y)$.

As one might expect, $P_{f^n} = P_f^n$ which means that the Frobenius-Perron operator for the n-fold composition of the transformation f is the same as the n-fold composition of the Frobenius-Perron operator for f. As such, the Frobenius-Perron operator is the discrete-time Fokker-Planck operator [38, 50] for the special case of a deterministic state equation.

A density function that is a fixed point of P, i.e., a density function $p(x)$ which satisfies $P_f(p(x)) = p(x)$, is called a *stationary density* of f since the density function induced by f is the same as the original density function. It follows directly from the defining equation for the Frobenius-Perron operator that if $p(x)$ is a stationary density of f, and μ_p denotes the measure induced by $p(x)$, i.e., $\mu_p(B) \equiv \int_B p(x)\, dx$ for each $B \in \beta$, then f is a measure-preserving transformation on the probability space (X, β, μ_p).

As discussed in Chapter 2, a measure-preserving transformation f (on (X, β, μ)) is *ergodic* if the only invariant, measurable sets (i.e, sets $B \in \beta$ satisfying $f^{-1}(B) = B$) have measure 0 or 1. In some references [50, 55], the definition of ergodicity does not require that f be measure-preserving, only that it be nonsingular. However, most properties generally associated with an ergodic transformation, such as the equivalence of time averages and ensemble averages, are valid only when the transformation is measure-preserving.

Our interest in ergodic transformations in this chapter differs from that in earlier chapters in which we exploited the ergodicity of dissipative, chaotic maps in a heuristic MMSE state estimator. Of interest in this chapter is the fact that an ergodic transformation has at most one stationary PDF of the Frobenius-Perron operator P_f. In particular, if a transformation f is measure-preserving and ergodic on (X, β, μ) and μ has a corresponding density function $p(x)$, then $p(x)$ is the unique stationary density function of P_f. We exploit this property of ergodic, measure-preserving transformations later in the chapter to synthesize signals which give rise to random variables with specified PDFs.

An equivalent condition for ergodicity of a measure-preserving transformation f is that the following condition hold for all sets $A, B \in \beta$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mu(A \cap f^{-i}(B)) = \mu(A)\,\mu(B) \tag{6.12}$$

which means that on average A and $f^{-i}(B)$ are independent of each other. A related but stronger property than ergodicity, not used in this chapter, is that of being *strong mixing.* A measure-preserving transformation f on the probability space (X, β, μ) is strong mixing if the following condition holds for all sets $A, B \in \beta$:

$$\lim_{n\to\infty} \mu(A \cap f^{-i}(B)) = \mu(A)\,\mu(B) \tag{6.13}$$

which means that A and $f^{-i}(B)$ asymptotically become independent of each other.

Finally, an even stronger concept than strong mixing is *exactness.* A measure-preserving transformation T on the probability space (X, β, μ) is exact if the following condition holds for all sets $B \in \beta$ with $\mu(B) > 0$:

$$\lim_{n\to\infty} \mu(f^n(B)) = 1 \tag{6.14}$$

Essentially, this condition means that the successive images under f of each measurable set, even those of arbitrarily small measure, expand until they cover almost all of X.

As discussed in [50], there is a strong relation among ergodicity, mixing, and exactness. In particular, an exact transformation is mixing, and a mixing transformation is ergodic. Therefore, if f is an exact transformation on (X, β, μ) where μ has corresponding density function $p(x)$, then $p(x)$ is the unique stationary density function of f. In addition, given

any initial density function $p_0(x)$ on X, $P_f^n(p_0(x))$ converges (in norm) as n goes to infinity to $p(x)$. Therefore, not only does an exact transformation have a unique stationary density function, but all density functions converge to this density function under successive applications of the Frobenius-Perron operator .

Although it is often difficult to determine whether a given system is ergodic, mixing, or exact, the definitions of these properties are straightforward and the properties exhibited by systems with these properties are well-understood. Furthermore, for the systems of interest in this chapter, in particular MC maps, the underlying measure space is extremely simple. This situation contrasts markedly with the one considered earlier in the thesis, where defining the nebulous concept of chaos and defining an appropriate topology or measure space on the attractors of dissipative systems suspected of being chaotic were formidable tasks.

We now consider the relation among the ergodic concepts defined above and the topological property—sensitive dependence on initial conditions—most often associated with chaos. For convenience, we consider the measure space (X, β, μ) with X denoting the unit interval, β denoting the Borel σ-algebra, and μ denoting Lebesgue measure; and we use the standard topology on the unit interval with basis given by the intersection of open intervals with the unit interval. Our first observation is that an ergodic system need not have sensitive dependence on initial conditions. For example, the *rotation map* with state equation given by

$$x(n+1) = f(x(n)) = x(n) + \alpha \pmod 1 \tag{6.15}$$

is ergodic if the constant α is irrational [87]. However, the map does not have sensitive dependence on initial conditions since distances between points on the unit interval are preserved by f, except for a wrap-around effect at the endpoints.

However, exactness of a map f implies sensitive dependence on initial conditions. This follows almost directly from the definition of exactness. Specifically, given any point x on the unit interval and any subinterval J_x containing x, $\lim_{n\to\infty} \mu(f^n(J_x)) = 1$ by definition of exactness. Therefore, for any $\delta \in [0,1]$, there exists an integer $N(x,\delta)$ such that $\mu(f^{N(x,\delta)}(J_x)) > 1 - \delta$. As such, there exists a point $y \in J_x$ for which $|f^{N(x,\delta)}(y) - f^{N(x,\delta)}(x)| > \frac{1-\delta}{4}$.

In contrast, a map is not necessarily exact even if it is both ergodic and has sensitive

dependence on initial conditions. The map $g(x)$ shown in Figure 6-4 (a) is ergodic and has sensitive dependence on initial conditions. We do not formally prove these claims here,

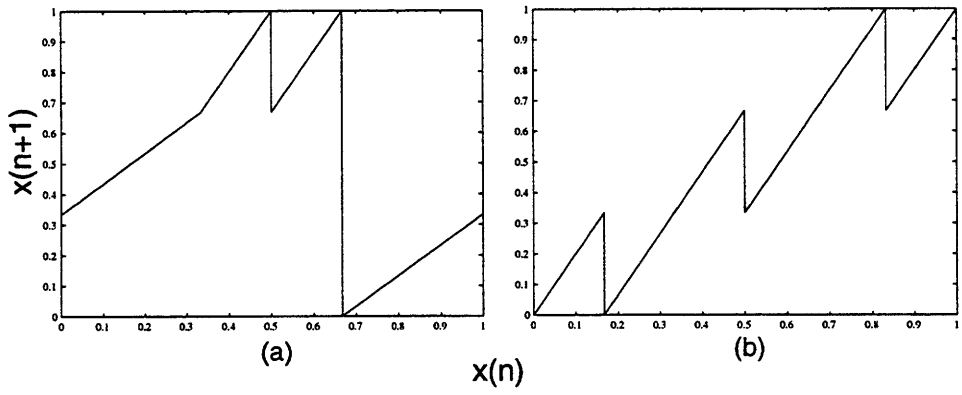

Figure 6-4: (a) Map g having sensitive dependence on initial conditions that is ergodic but not exact; (b) g^3

but sketch one approach for proving them. One can verify both claims by considering the composed map g^3 shown in Figure 6-4 (b). As indicated by the figure, g^3 consists of three separate shift maps, one acting on each of the subintervals [0,1/3[, [1/3,2/3[, and [2/3,1]. Each of these subintervals is invariant under g^3. Ergodicity is established by using the facts that the shift map is ergodic (as shown in [50, 87]) and that g applied to each of the three subintervals [0,1/3[, [1/3,2/3[, and [2/3,1] consists of either a translation or a shift followed by a translation. With these facts, it is straightforward to verify the equivalent condition for ergodicity proven in [87], which requires that for each measurable set A with nonzero measure the following condition holds:

$$\mu\left(\bigcup_{n=1}^{\infty} f^{-n}(A)\right) = 1. \tag{6.16}$$

Sensitive dependence on initial conditions follows from the fact that the shift map exhibits sensitive dependence on initial conditions and g^3 consists of three shift maps, one applied to each of the three invariant sets of g^3 given by [0,1/3[, [1/3,2/3[, and [2/3,1].

Therefore, the map shown in Figure 6-4 (a) is ergodic and has sensitive dependence on initial conditions. However, the map is not exact as it is straightforward to verify that $\lim_{n\to\infty} \mu(f^n(J)) = 1/3 \neq 1$ for every subinterval $J \subset [0, 1/3[$.

Concepts similar to ergodicity and exactness apply to the transition probability matrices (TPMs) of homogeneous, finite-state Markov chains. In particular, let $P = [p_{ij}]$ denote the

TPM of a homogeneous, finite-state Markov chain, where p_{ij} denotes the probability of transitioning from state i to state j in one time step, and let $P^n = [p_{ij}^n]$ denote the n step TPM of the Markov chain, where p_{ij}^n denotes the probability of transitioning from state i to state j in exactly n time steps. P^n is the n^{th} power of P, but in general p_{ij}^n is not the n^{th} power of p_{ij}. The TPM is *irreducible* if for each pair of indices (i, j), there exists a positive integer n_{ij} such that $p_{ij}^{n_{ij}} > 0$. Intuitively, irreducibility of the transition matrix means that it is possible to eventually get from each state of the Markov chain to every other state. A well-known result from linear algebra [29, 49] is that every irreducible matrix has a unique, invariant left eigenvector. It follows from this that every Markov chain with an irreducible TPM has a unique, invariant state probability vector, i.e., a row vector of probabilities $\boldsymbol{\Pi}$ that satisfies $\boldsymbol{\Pi} P = \boldsymbol{\Pi}$. A Markov chain with irreducible TPM is also ergodic with respect to this invariant probability vector.

If in addition to being irreducible, the TPM has the property that there is a single, positive integer m such that $p_{ij}^m > 0$ for all pairs of indices (i, j), then the TPM is also *primitive*. Intuitively, a Markov chain with primitive TPM is one in which it is possible to get from each state to itself and to each other state in exactly the same number of time steps. A Markov chain with primitive TPM has the property that any initial state probability vector converges to the unique, invariant probability vector as time goes to infinity.

6.5.2 EMC Maps, Markov Chains, and Stationary PDFs[1]

Unless indicated otherwise, in this section all EMC maps are assumed to be maps of the unit interval to itself. By appropriately merging results from [12, 25, 55, 87], we can establish the following, strong connection between EMC maps and the Markov chains which they give rise to.

Proposition 3: a. Ergodicity of an EMC map for which each subinterval of the unit interval has nonzero measure is equivalent to irreducibility of the TPMs of the Markov chains the map gives rise to. Thus, an ergodic EMC map has a unique stationary PDF, and each Markov chain it gives rise to has a unique, invariant state probability vector.

b. Exactness of an EMC map for which each subinterval of the unit interval has nonzero

[1]Some of the work in this section was performed in conjunction with S. Isabelle at MIT

measure is equivalent to primitivity of the TPMs of the Markov chains the map gives rise to. Thus, not only does an exact EMC map have a unique stationary PDF, but given any nonsingular PDF for $x(0)$, the PDF of $x(n)$ converges as n goes to infinity to the unique stationary PDF. Similarly, each Markov chain which the map gives rise to has a unique, invariant state probability vector which all other initial state probability vectors converge to.

Proof: (See Appendix B)

As noted in [12] and used in the proof of Proposition 3, the stationary PDF of an ergodic EMC map is piecewise constant. A simple way to find it is the following. First, find a Markov partition $\{I_j\}_{j=1}^n$ and the TPM of the corresponding Markov chain. Next, find the invariant probability vector for this TPM. Let $\boldsymbol{\Pi} = [\pi_1, \cdots, \pi_n]$ denote this vector, where π_j denotes the invariant probability associated with the state corresponding to subinterval I_j. The stationary PDF for all points in I_j is given by $\pi_j/\lambda(I_j)$.

Although EMC maps which give rise to Markov chains with irreducible TPMs have piecewise constant stationary PDFs, one can not independently specify this PDF and a specific Markov chain which the map gives rise to. However, using an EMC map as a building block, one can synthesize a Markov map (which may or may not be an EMC map) that has any specified stationary PDF and which also gives rise to a Markov chain with any specified TPM. To do so, one utilizes the following relation, established in [12], among maps with stationary PDFs and maps derived from them via homeomorphisms:

Proposition:[12] For any EMC map f which gives rise to a Markov chain with irreducible TPM and any differentiable homeomorphism $h : I \rightarrow I$, the transformation $g \equiv h \circ f \circ h^{-1}$ has a unique stationary PDF $p_G(x)$, which is given by $p_G(x) = \frac{p_F(h^{-1}(x))}{|h'(h^{-1}(x))|}$, where $p_F(x)$ is the unique stationary PDF of f and $h'(h^{-1}(x))$ denotes the derivative of h evaluated at $h^{-1}(x)$.

The above proposition is essentially a restatement of the relation considered earlier (in the discussion of the Frobenius-Perron operator) between the PDF of a random variable and the PDF resulting from a memoryless transformation on that random variable. In this case, the random variable is $x(n)$ with PDF p_F, h is the transformation, and p_G is the PDF of the transformed random variable $h(x(n))$.

As noted in [68, p. 261] and in other probability textbooks, given a random variable v

with PDF p_V and distribution function F_V where

$$F_V(v_0) \equiv \int_{-\infty}^{v_0} p_V(v)\, dv \tag{6.17}$$

one can create a random variable w with specified PDF p_W or specified distribution function $F_W(w_0) = \int_{-\infty}^{w_0} p_W(w)\, dw$ using the transformation $F_W^{-1}(F_V)$, so that

$$w = F_W^{-1}(F_V(v)). \tag{6.18}$$

An implicit assumption with this transformation is the invertibility of F_W, which is always the case if $p_W(w) > 0$ for all real w or if the domain of F_W is restricted to those w for which $p_W(w) > 0$.

We can use this result for transforming PDFs and the earlier proposition to synthesize a Markov map that has any specified (well-behaved) stationary PDF and which gives rise to a Markov chain with any specified irreducible TPM. The procedure is as follows. First, synthesize an EMC map f which gives rise to the desired Markov chain. From Proposition 3, we know that such a map has a stationary PDF that is piecewise constant and in particular constant over each element of any Markov partition. Furthermore, the PDF must be nonzero almost everywhere on the unit interval. Otherwise, there would be a subinterval of zero density, which would necessarily correspond to a partition element or union of partition elements in any Markov partition, since the stationary PDF is constant over partition elements. Consequently, this subinterval of zero density would be associated with at least one state in the associated Markov chain. The invariant probability for this state would necessarily be zero, since its value is given by the integral of the stationary PDF over the corresponding subinterval. However, each element of the invariant vector of an irreducible matrix must be nonzero, by convention. Thus, the stationary PDF must be nonzero almost everywhere.

Let p_F denote the stationary PDF of f and F_F the corresponding distribution function. Let p_G denote the specified, stationary PDF and F_G the corresponding distribution function. Now let $h = F_G^{-1}(F_F)$ in the above proposition. If h is a differentiable homeomorphism, then by this proposition and the result for transforming PDFs, g will have the desired stationary PDF. Note however that the requirement that h be a differentiable homeomorphism requires that the desired PDF p_G be sufficiently well-behaved in a mathematical sense.

In addition, g gives rise to the same Markov chains as f. An informal argument as to why this is true is the following. Since h is a homeomorphism, it is by definition continuous and invertible. Because of the continuity of h, each subinterval I_j in a Markov partition $M_f = \{I_j\}$ of f is mapped by h to a subinterval. Because of the invertibility of h, the image under h of any two different partition elements do not intersect. Therefore, the set of subintervals $M_g = \{h(I_j)\}$ forms a partition for g. Also, since h is a homeomorphism, the partition is a Markov partition since g acting on each element in M_g is equivalent to f acting on the corresponding elment in M_f. That is, if $f(I_j) = I_k \cup I_l$, then $g(h(I_j)) = h(I_k) \cup h(I_l)$ which is easily verified:

$$\begin{aligned} g(h(I_j)) &\equiv (h \circ f \circ h^{-1})(h(I_j)) && (6.19) \\ &= h(f(h(h^{-1}(I_j)))) = h(f(I_j)) && (6.20) \\ &= h(I_k \cup I_l) = h(I_k) \cup h(I_l). && (6.21) \end{aligned}$$

If $\{S_i\}$ denote the states associated with the partition M_f such that S_i is associated with I_i, and if we associate state S_i with $h(I_i)$ as well, then state sequences that arise under the dynamics of g are sample paths of a Markov chain, the same Markov chain as for f. This follows from two facts. First, $\mu_f(I_j) = \mu_g(h(I_j))$ where μ_f and μ_g are the measures induced by p_F and p_G respectively. This is a consequence of p_G being the density induced by the transformation h. Second, the transition probabilities are the same. This is a consequence of the following relation which we verify below:

$$\frac{\mu_f(I_{jk})}{\mu_f(I_j)} = \frac{\mu_g(h(I_j)_k)}{\mu_g(h(I_j))} \tag{6.22}$$

where I_{jk} denotes the portion of I_j mapped to I_k by f and $h(I_j)_k$ denotes the portion of $h(I_j)$ mapped to $h(I_k)$ by g. The ratio $\frac{\mu_f(I_{jk})}{\mu_f(I_j)}$ is the transition probability from state S_j to S_k for map f and $\frac{\mu_g(h(I_j)_k)}{\mu_g(h(I_j))}$ is the transition probability from state S_j to S_k for map g. This is a generalization (briefly discussed in [61]) of the earlier result for piecewise linear maps of the unit interval in which transition probabilities were given by the ratio of interval lengths. For the EMC map f, the definitions of transition probabilities are identical, as they must be since p_F is constant over partition elements.

To verify (6.22), we must show that $h(I_{jk}) = h(I_j)_k$. To do so, we first note the following

relations which are straightforward to verify:

$$I_{jk} = I_j \cap f^{-1}(I_k) \tag{6.23}$$

$$h(I_j)_k = h(I_j) \cap g^{-1}(h(I_k)) \tag{6.24}$$

$$g^{-1} = h \circ f^{-1} \circ h^{-1}. \tag{6.25}$$

Substituting (6.25) in (6.24) yields

$$h(I_j)_k = h(I_j) \cap (h \circ f^{-1} \circ h^{-1})(h(I_k)) \tag{6.26}$$

$$= h(I_j) \cap h(f^{-1}(I_k)) \tag{6.27}$$

$$= h(I_j \cap f^{-1}(I_k)) \tag{6.28}$$

$$= h(I_{jk}) \tag{6.29}$$

where the next to last equality holds because h is a homeomorphism.

The Markov partitions of g are uniquely determined by those of f and the transformation h. Whereas the map f might give rise to uniform Markov partitions, in general this is not true of g. For example, if the region of support of g is all of R (e.g., p_G is a Gaussian PDF), then two partition elements will have infinite length for any Markov partition, whereas all others will have finite length.

The question arises as to the relevance of unique, stationary PDFs for a map. The answer is that such a PDF, if it exists, determines the concentration of orbit points on the real line, for orbits generated by most initial conditions. For example, Figures 6-5, 6-6, and 6-7 depict EMC maps and typical orbit segments generated by each map. The maps shown in Figures 6-6 (a) and 6-7 (a) were both derived from the map in Figure 6-5 (a) using the procedure outlined above. The stationary PDF of the map shown in Figure 6-5 (a) is constant over the unit interval; the stationary PDF of the map shown in Figure 6-6 (a) equals 4 over the subinterval $[.4, .6[$ and .25 elsewhere; and the stationary PDF of the map shown in Figure 6-7 (a) equals 4 over the subinterval $[.0, .2[$ and .25 elsewhere. For each map, the orbit segments reflect the stationary PDFs, with orbit points concentrated in regions where the probability density is large in Figures 6-6 (b) and 6-7 (b). The relation between stationary PDFs and the behavior of orbits coupled with the ability to independently specify the stationary PDF of a Markov map and the Markov chains it gives rise to may have practical value, such as

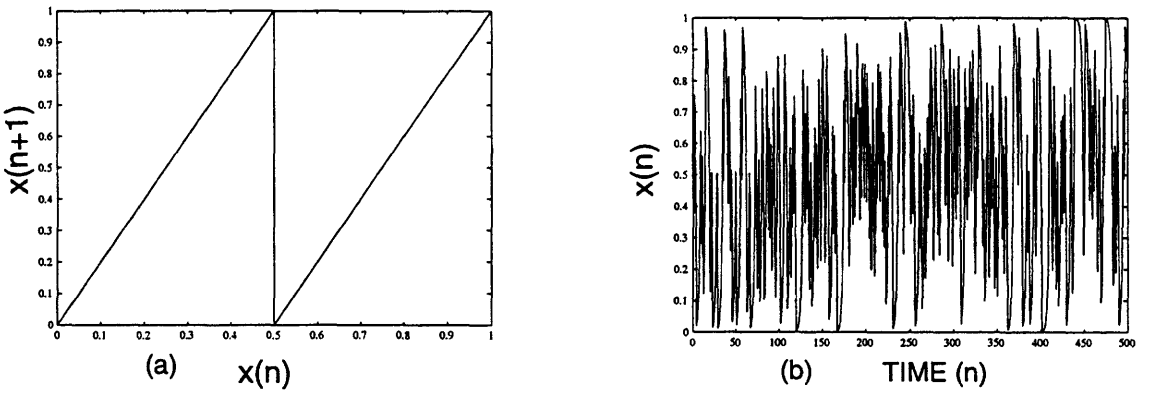

Figure 6-5: (a) EMC map; (b) Typical Orbit segment

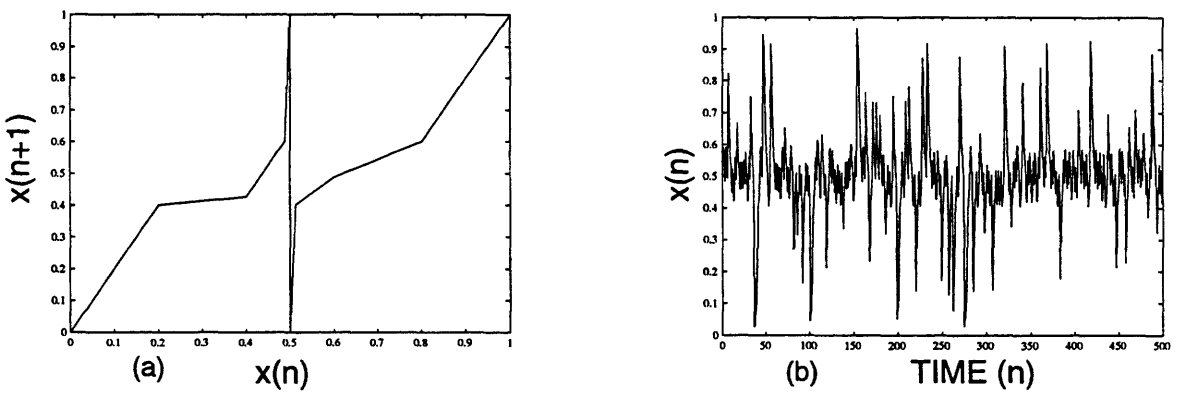

Figure 6-6: (a) EMC map; (b) Typical Orbit segment

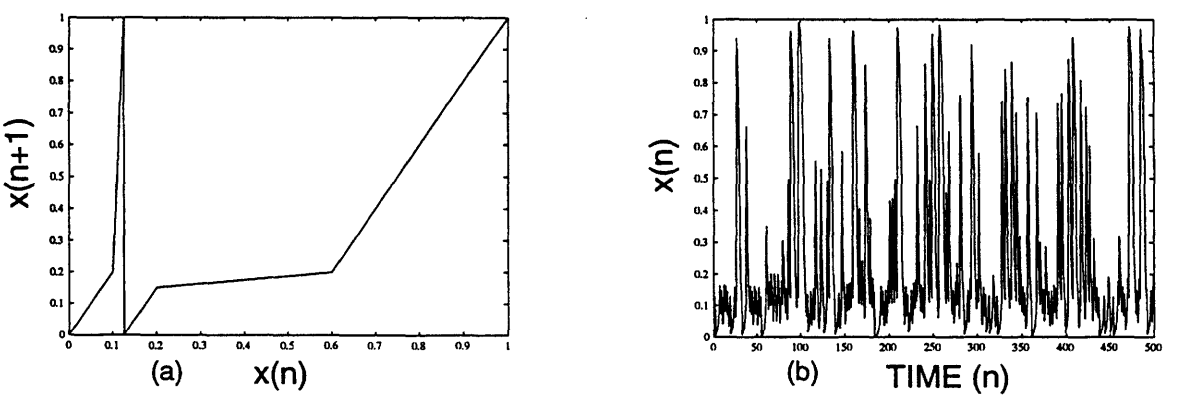

Figure 6-7: (a) EMC map; (b) Typical Orbit segment

in the design of signals for communication applications.

6.6 Multidimensional Markov Maps

It is straightforward to synthesize multidimensional Markov maps that give rise to Markov chains, a surprising fact in light of the difficulty generally encountered in finding Markov partitions for maps with dimension greater than one. With one-dimensional maps, partition elements of Markov partitions are subintervals or one-dimensional volume elements, and the requirement for a partition to be a Markov partition is that each subinterval in the partition maps to a union of subintervals in the partition. Similarly, with m-dimensional maps, partition elements for Markov partitions are m-dimensional volume elements, and the requirement for a partition to be a Markov partition is that each volume element in the partition maps to a union of volume elements in the partition. As one might expect, synthesizing maps of the m-dimensional unit cube for which one can find volume elements satisfying this requirement is in general a daunting task.

One set of piecewise linear, m-dimensional maps for which one can find Markov partitions and which also give rise to Markov chains is the set of hyperbolic toral automorphisms [20]. However, although a Markov partition is known to exist for a given hyperbolic toral automorphism, finding the partition elements is a fairly complex procedure even in two dimensions [3, 10, 11].

We now discuss techniques for synthesizing m-dimensional MC maps which give rise to Markov chains by using one-dimensional MC maps as building blocks. The simplest technique is to start with m one-dimensional MC maps and use each, or more precisely the state of each, as a component in an m-dimensional state vector. For example, if $x_i(n+1) = f_i(x_i(n))$ denotes the state equation for the i^{th} one-dimensional MC map, then the state equation for a two-dimensional MC map is given by

$$\boldsymbol{x}(n+1) = \begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \mathbf{F}(\boldsymbol{x}(n)) = \begin{bmatrix} f_1(x_1(n)) \\ f_2(x_2(n)) \end{bmatrix}. \tag{6.30}$$

It is straightforward to show that if a Markov partition for f_1 has N_1 elements and a Markov partition for f_2 has N_2 elements, then the corresponding Markov partition for $\boldsymbol{F}$ has $N_1 N_2$ elements, with each partition element a rectangle in the unit square with

each side parallel to one of the coordinate axes. Furthermore, Markov partitions for the multidimensional map also give rise to Markov chains. It is straightforward to show that if I_{ij} denotes the partition element formed from the i^{th} partition element of f_1 and the j^{th} partition element of f_2, then the transition probability from I_{ij} to I_{kl} is given by $p_{ik}\, q_{jl}$ where p_{ik} is the transition probability from the i^{th} state to the k^{th} state of f_1 and q_{jl} is the transition probability from the j^{th} state to the l^{th} state of f_2.

The dynamics of multidimensional MC maps synthesized this way are rather trivial since the dynamics along each component are independent and are those of a one-dimensional MC map. However, by relaxing the restriction that the multidimensional MC map be a mapping from the unit m-cube to itself, one can easily extend this synthesis approach to create multidimensional MC maps with nontrivial dynamics. To do so, one applies a similarity transformation to the original multidimensional transformation. For example, let $\boldsymbol{A}$ denote an invertible, m-dimensional matrix and let $\boldsymbol{x}(n+1) = \boldsymbol{F}(\boldsymbol{x}(n))$ denote the state equation of an m-dimensional MC map synthesized by using a one-dimensional MC map for each component. Then, $\boldsymbol{y}(n) = \boldsymbol{A}\boldsymbol{x}(n)$ is also an MC map with state equation $\boldsymbol{y}(n+1) = \boldsymbol{G}(\boldsymbol{y}(n))$ given by

$$\boldsymbol{y}(n+1) = \boldsymbol{G}(\boldsymbol{y}(n)) = \boldsymbol{A}\boldsymbol{F}\boldsymbol{A}^{-1}\boldsymbol{y}(n). \tag{6.31}$$

An informal argument as to why $\boldsymbol{G}$ is an MC map is the following. Because $\boldsymbol{A}$ is invertible, it is a homeomorphism. Therefore, one can use the same reasoning as used in the earlier discussion about synthesizing Markov maps with specified PDFs to argue that if $M_F = \{I_{ij}\}$ is a Markov partition for $\boldsymbol{F}$ then $\{\boldsymbol{A}\, I_{ij}\}$ is a Markov partition for $\boldsymbol{G}$. Specifically, because $\boldsymbol{A}$ is continuous, each element of M_F is mapped by $\boldsymbol{A}$ to a connected region (a parallelogram since A is linear) and because $\boldsymbol{A}$ is invertible and the elements of M_F are disjoint, the images of these elements are disjoint as well. Furthermore, the fraction of each partition element $\boldsymbol{A}\, I_{ij}$ mapped to a partition element $\boldsymbol{A}\, I_{kl}$ by $\boldsymbol{G}$ is the same as the fraction of I_{ij} mapped to I_{kl} by $\boldsymbol{F}$, because $\boldsymbol{A}$ is a homeomorphism. Therefore, the state transition probabilities are the same for the two maps.

As noted above, since each partition element I_{ij} is a rectangle with each side parallel to a coordinate axis, it follows that each partition element $\boldsymbol{A}\, I_{ij}$ is a parallelogram and not necessarily a rectangle. Also, in contrast to the components of the state vector of $\boldsymbol{F}$ for

which each is the state of a one-dimensional MC map with dynamics independent of the other components, the components of the state vector of $\boldsymbol{G}$ are not independent in general, and each does not correspond to the state of a one-dimensional MC map.

Figure 6-8 depicts 4000 consecutive orbit points generated by a two-dimensional MC map $\boldsymbol{F}$ synthesized by using the maps shown in Figures 6-6 (a) and 6-7 (a) to generate each component of the state vector. One component of the state vector is plotted versus the other in

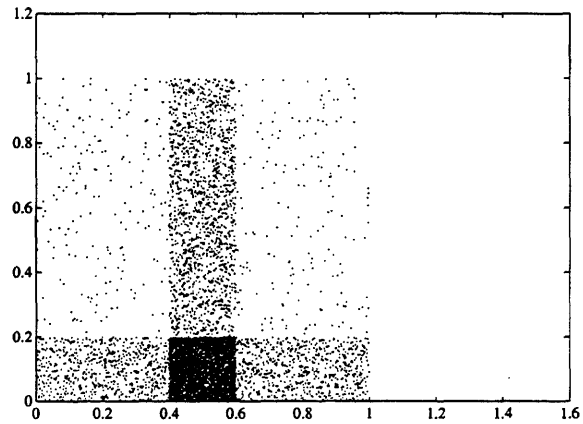

Figure 6-8: Two-dimensional EMC map

Figure 6-8. Figure 6-9 depicts 4000 consecutive orbit points generated by the map $\boldsymbol{AFA}^{-1}$ where $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}. \tag{6.32}$$

By appropriately selecting the one-dimensional maps for the components of $\boldsymbol{F}$ and the transformation $\boldsymbol{A}$, one can generate multidimensional MC maps with interesting attractor patterns.

We can generalize this result by letting $\boldsymbol{y}(n) = \boldsymbol{H}(\boldsymbol{x}(n))$ for any homeomorphism $\boldsymbol{H}$, which leads to the map with state equation $\boldsymbol{y}(n+1) = \boldsymbol{G}(\boldsymbol{y}(n))$ given by

$$\boldsymbol{y}(n+1) = \boldsymbol{G}(\boldsymbol{y}(n)) = \boldsymbol{H}(\boldsymbol{F}(\boldsymbol{H}^{-1}(\boldsymbol{y}(n)))). \tag{6.33}$$

Using similar reasoning as used in the preceding section, one can show that because $\boldsymbol{H}$ is a homeomorphism, the image of each (rectangular) partition element I_{ij} under $\boldsymbol{H}$ is a connected region in the range of $\boldsymbol{H}$, the image is disjoint from the images of all other

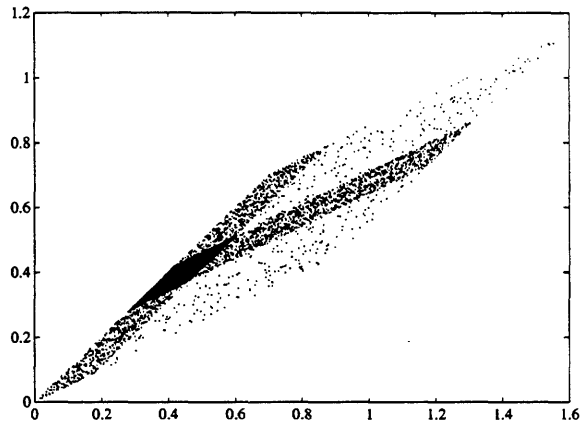

Figure 6-9: Two-dimensional EMC map derived from map shown in Figure 6-8

partition elements, and the set $\{\boldsymbol{H}(I_{ij})\}$ is a Markov partition for $\boldsymbol{G}$. One can also show that the two maps give rise to the same Markov chains.

6.7 Practical Considerations

The discussion thus far has ignored a practical problem one encounters when simulating certain MC maps on digital computers. The problem is due to the ability of digital computers to store and manipulate only dyadic rationals. As noted earlier, every rational number on the unit interval is an eventually periodic point of any MC map for which the slope of each affine segment is integer-valued. As such, each initial condition is either periodic or eventually periodic, and thus no orbits are dense, exhibiting the randomlike behavior expected of orbits generated by such maps.

Although one might argue that this situation arises with all maps, not just MC maps, the effects are more pronounced with certain MC maps. For example, consider the shift map given by $x_{n+1} = 2x_n \pmod 1$. The name *shift map* arises from the fact that the map shifts the binary representation of x_n left one place and retains the fractional part of the result. All rational points on the unit interval are periodic or eventually periodic points of the shift map, and in theory all irrational points have dense orbits. However, since a computer stores only dyadic rationals and since the shift map shifts the binary representation of x_n left one place, it follows that the computer-generated orbit for the shift map with any initial condition x_0 becomes zero-valued after a finite amount of time.

Figure 6-10 depicts the computer-generated orbit corresponding to the initial condition $\frac{\sqrt{2}}{2}$. Despite the fact the actual initial condition is irrational, the computer approximates

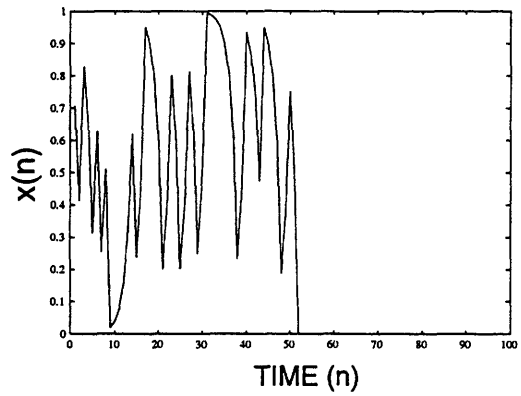

Figure 6-10: Orbit segment for shift map with initial condition $\frac{\sqrt{2}}{2}$.

the initial condition with a dyadic rational, and the orbit becomes zero-valued after a finite amount of time.

There are several practical, though not totally theoretically justifiable ways to circumvent this undesirable situation when simulating MC maps such as the shift map on a computer. First, for each affine segment with slope that is integer-valued and even, one can perturb the slope slightly. For example, one might replace a slope of 2 with a slope of 1.9999999999. Unfortunately, the substitution changes the dynamics of the map as well as any stationary PDFs. In fact, the resulting map may not even have a stationary PDF even though the original map is ergodic. Intuitively, one would expect the dynamics of the new map to be close to those of the original. However, it is difficult if not impossible to analytically evaluate the effect of perturbed slopes, and it is unclear if the invariant density or dynamics of the perturbed map converge to those of the unperturbed map as the perturbations go to zero.

Alternatively, one can add a small driving noise term to the state equation of an MC map. This yields the following non-deterministic state equation:

$$x_{n+1} = f(x_n) + w_n \tag{6.34}$$

where $\{w_n\}$ is a white-noise sequence. If the PDF of each noise term w_n is constant with region of support over $[-\alpha, \alpha]$ for some constant α, then as shown in [13], the stationary PDF of the driven system converges in L^1-norm to that of the undriven system as the

driving noise sequences converges to an impulse at the origin, i.e., as the bounding constant α goes to zero. Furthermore, in contrast to the undriven system which may have an infinite set of periodic points, the driven system has no periodic points and as a result the orbit generated by each initial condition exhibits the expected randomlike behavior.

Similarly, if one expresses w_n as $\epsilon\, e_n$ where ϵ is a small positive constant and $\{e_n\}$ is an independent, identically distributed random sequence with nonzero density over R, then the stationary PDF of the driven system converges to that of the undriven system in L^1-norm as $\epsilon \rightarrow 0$, as a consequence of a more general result given by [50, p.289:Theorem 10.6.1].

A third technique, the one used for the examples in this chapter, consists of scaling and then rescaling each of the affine mappings comprising the MC map. In particular, the affine map

$$f(x) = \tau\, x + \beta \tag{6.35}$$

is replaced by the theoretically equivalent affine map

$$f(x) = \frac{1}{\gamma}[(\gamma\,\tau)\, x + (\gamma\,\beta)] \tag{6.36}$$

where γ is an irrational number and the parenthesized expressions are evaluated first. Figures 6-11 (a) and (b) show the first 100 orbit points and orbit points 1001 to 1100, respectively, for the same map and initial condition used for Figure 6-10 but with the affine transformations comprising the map scaled and rescaled as indicated above by the constant $\gamma = \sqrt{11}$. The orbit exhibits the randomlike behavior expected with this map and initial

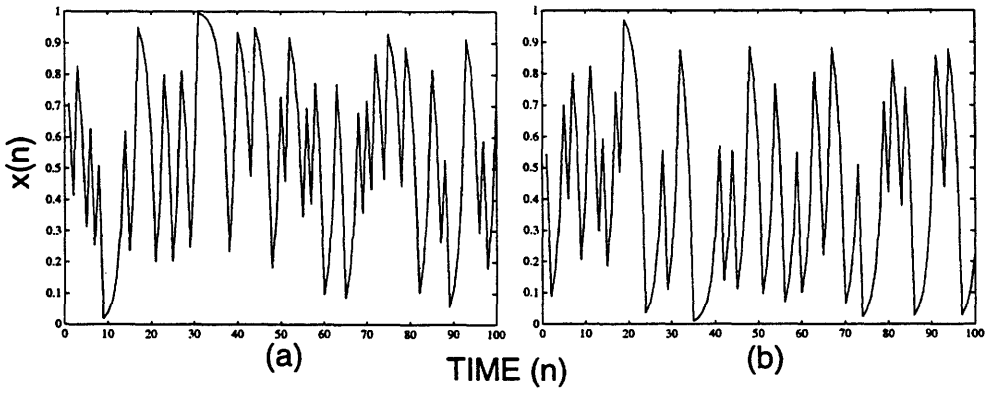

Figure 6-11: Orbit segments for shift map with irrational initial condition and affine transformation scaled and rescaled by $\gamma = \sqrt{11}$. (a) Orbit points 1–100: (b) Orbit points 1001–1100.

condition.

An interesting property of this third technique is that orbit segments of EMC maps generated using it, or at least those EMC maps for which the slope of each affine segment is an integer with absolute value greater than one, are deterministic both in a theoretical and in a certain practical sense. In particular, given an EMC map f and an $(N+1)$-point orbit segment $\{x(i) = f^i(x(0))\}_{i=0}^N$ generated with this technique, the *backward orbit segment* $\{y(i)\}_{i=0}^N$ obtained by defining $y(N) = x(N)$ and $y(n) = f_n^{-1}(y(n+1))$ for $0 \leq n < N-1$, where f_n denotes the affine mapping associated with $x(n)$, is the same as $\{x(i)\}_{i=0}^N$ to within machine precision. In other words, by running the final orbit point $x(N)$ through successive compositions of the inverse system f_n^{-1} implicitly defined by the orbit segment, one recovers the orbit segment to within machine precision. Figure 6-12 depicts the average, point-by-point, squared reconstruction error normalized by the variance of the original orbit segment $\{x(i)\}_{i=0}^N$ of the shift map, with the original orbit segment generated using the third technique discussed above. Figure 6-13 (b) shows analogous results for the EMC map shown

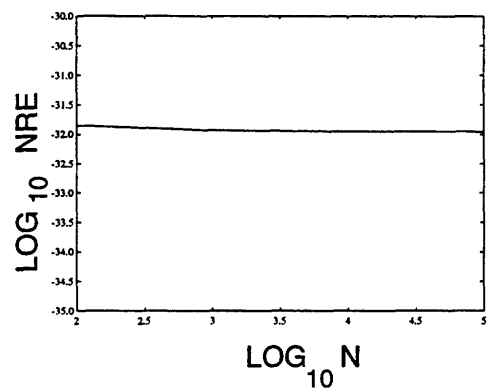

Figure 6-12: Normalized, average, point-by-point, squared reconstruction error (NRE) with orbit segments of length $N+1$ for shift map.

in Figure 6-13 (a).

As suggested by the figures, in contrast to orbit segments of the dissipative maps considered in earlier chapters, an orbit segment of an EMC map is recoverable from the final orbit point if the sequence of affine segments which gave rise to the orbit segment is known. This invertibility or recoverability property of EMC maps with affine segments having integer-valued slopes is due in part to the contractive nature of the inverse system f_n^{-1}. In the next chapter, we exploit this property in the derivation and implementation of an ML state estimator for use with these maps.

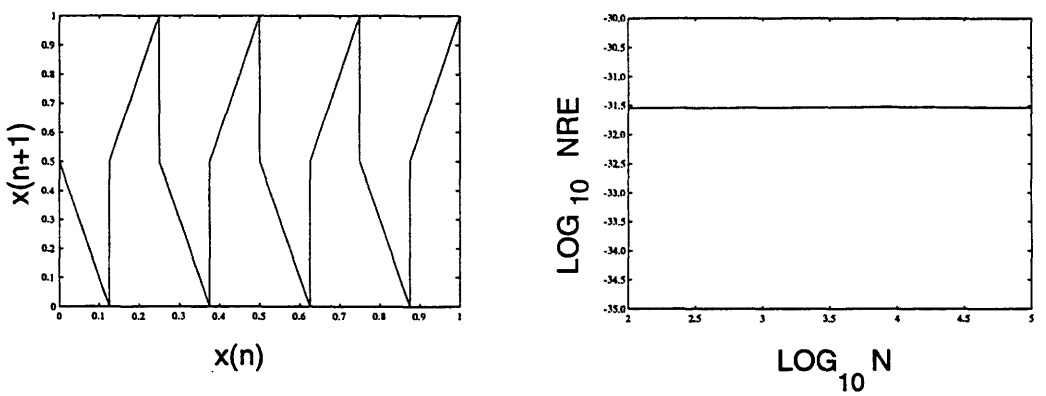

Figure 6-13: EMC map and normalized, average, point-by-point, squared reconstruction error (NRE) with orbit segments of length $N + 1$. (a) EMC map; (b) NRE.

Chapter 7

Detection and State Estimation with MC Maps

7.1 Introduction

As in Chapter 6, the class of maps of interest in this chapter is the class of MC maps, a class of piecewise linear unit-interval maps for which the members give rise to Markov chains. However, in contrast to Chapter 6 which focused on properties of these maps and synthesis of signals using these maps as building blocks, this chapter focuses on the exploitation of these properties in practical, optimal and suboptimal detection and state estimation algorithms for use with these maps. We first consider detection, in particular the problem of discriminating among a finite number of EMC maps when given a noise-corrupted orbit segment generated by one of the maps. We introduce a hidden-Markov-model (HMM) representation of this discrimination problem, with the representation being an exact one if the orbit points are properly quantized and an approximate one otherwise. We then exploit this representation in computationally efficient, optimal and suboptimal discrimination algorithms, and we assess the performance of the algorithms. We also exploit this HMM representation in iterative, optimal and suboptimal ML estimation algorithms for estimating both an unknown, constant scale factor applied to the orbit points as well as the variance of the corrupting noise. We conclude our discussion of detection by briefly considering an experimentally determined discriminability metric for use with these maps.

Our focus then shifts to state estimation with MC maps, in particular the problem of

obtaining the ML estimate of each point of an orbit segment generated by an MC map when given a noise-corrupted version of that orbit segment. We first consider the issues and complications that arise when attempting to perform ML state estimation with MC maps. Next, we introduce an HMM representation of the state estimation problem. The representation is not exact; but it constitutes a fundamental component of an ML state estimator which we derive and show to be optimal if the representation is appropriately chosen and suboptimal, but nonetheless potentially effective, otherwise.

7.2 Detection of EMC Maps

7.2.1 Problem Scenario

The close relation between EMC maps and Markov chains facilitates the detection or discrimination of these maps based on noise-corrupted orbit segments generated by the maps. In this section, we show that when the orbit points are properly quantized, one can perform optimal, computationally efficient discrimination among these maps, and when the outputs are not quantized, one can perform computationally efficient, albeit suboptimal discrimination.

The underlying problem scenario we consider is the following. We are given M one-dimensional EMC maps $\{f_i\}_{i=1}^{M}$, and an unobserved $(N+1)$-point orbit segment $X \equiv \{x(i)\}_{i=0}^{N}$ is generated by one of the maps. We are also given a set of $N+1$ observations $Y = \{y(i)\}_{i=0}^{N}$, where $y(i)$ is given by

$$y(n) = h_k(x(n)) + v(n), \quad 0 \leq n < N. \tag{7.1}$$

In this equation, $\{v(i)\}_{i=0}^{N}$ is a white-noise sequence which is assumed to be independent of the initial condition $x(0)$ and the chosen map f_k and for which the variance of each $v(i)$ is σ^2. Also, h_k is a memoryless transformation which may be dependent upon the map f_k. In Section 7.2.2 we consider the case in which h_k is a quantizer. In Section 7.2.3 we focus on the case in which h_k is the identity operator but briefly consider more general choices of h_k as well. In Section 7.4, we consider ML estimation of h_k when it is an unknown, constant, scale factor.

This scenario gives rise to several related problems. The problem of interest here is an M-

ary hypothesis testing problem in which one seeks to determine, ideally in an optimal way, which of the M maps generated the unobserved orbit segment X and thus gave rise to the observations Y. One motivation for considering this problem is its relevance to a potential scheme for secure M-ary communication. With this scheme, an EMC map is associated with each of the M signals. To transmit a signal, one generates a sufficiently long orbit segment X from the corresponding map and transmits this segment or a transformation h_k of the segment. If the channel is an independent, additive noise channel, the task at the receiver involves M-ary hypothesis testing, that is, determining which of the M maps generated the received, noise-corrupted orbit segment. This potential communication scheme has some similarity with those techniques for spread-spectrum communication which utilize binary chipping or pseudo-random noise (PN) sequences. However, in contrast to these techniques, the scheme does not require the receiver to be precisely synchronized to the transmitter. Only a sufficiently long subsegment of each transmitted orbit segment needs to be isolated for detection. Also, because of the flexibility in designing EMC maps, one could choose the M maps as a function of the expected noise level of the channel. At low noise levels, the maps and the associated signals could be chosen to be nearly indistinguishable so as to minimize the possibility of interception, whereas at high noise levels the maps could be chosen to be quite dissimilar, possibly *antipodal* in the binary case as discussed in [65, 66, 67].

A fundamental result from estimation theory is that for equally likely maps, the optimal detection rule, in a minimum probability of error sense, is to choose that map among the M, with the largest likelihood $p(Y|f_k)$, where $p(Y|f_k)$ is the PDF of the observation set Y conditioned on the map f_k having generated the orbit segment X giving rise to Y. The next two sections introduce optimal and suboptimal algorithms for calculating these likelihoods. In both sections, the transformations $\{h_k\}$, the M maps, and the variance σ^2 of the observation noise are assumed to be known. Section 7.4 considers optimal and suboptimal methods to partially overcome the need for these assumptions.

7.2.2 Detection with Quantized Orbit Points

One can efficiently compute the exact likelihoods used in the optimal detection rule when each of the M transformations h_k is a quantizer which associates a single, unique value with each element of a Markov partition for f_k. That is, if I_j^k denotes the j^{th} partition element for map f_k and H_j^k denotes the value associated with this partition element by h_k,

then $h_k(x) = H_j^k$ for all $x \in I_j^k$. (For example, H_j^k might be the midpoint or an endpoint of I_j^k). With the M quantizers chosen this way, the detection problem reduces to that of discriminating among M hidden Markov models (HMMs). As such, one can use the forward portion of the computationally efficient *forward-backward* algorithm [74, 75] to calculate the the M likelihoods $\{p(Y|f_k)\}_{k=1}^M$.

Specifically, for each map f_k the partition elements $\{I_j^k\}_{j=1}^{T_k}$ correspond to the *unobserved* states in the HMM associated with that map, where T_k denotes the number of partition elements for the Markov partition associated with f_k. For the problem scenario introduced earlier, $o_j^k(n)$, the output at time n associated with I_j^k, is given by

$$o_j^k(n) = H_j^k + v(n). \tag{7.2}$$

In other words, if f_k generated the orbit segment $X = \{x(i)\}_{i=0}^N$ and $x(n) \in I_j^k$, then $y(n)$, the observation at time n, is given by

$$y(n) = h_k(x(n)) + v(n) = o_j^k(n) = H_j^k + v(n). \tag{7.3}$$

With this observation equation, $p(y(n)|[x(n) \in I_j^k, f_k])$, the PDF of the observation $y(n)$ conditioned on the map f_k having generated the orbit segment and $x(n)$ being in partition element I_j^k, is given by

$$p(y(n)|[x(n) \in I_j^k, f_k]) = p_V\left(v(n) = y(n) - H_j^k\right) \tag{7.4}$$

where p_V is the PDF of each term in the white-noise sequence.

We define the *forward variable* $\alpha_j^k(n)$ as follows:

$$\alpha_j^k(n) = p(y(0), y(1), \cdots, y(n), x(n) \in I_j^k | f_k). \tag{7.5}$$

That is, $\alpha_j^k(n)$ is the joint PDF of the observations through time n and $x(n) \in I_j^k$, conditioned on map f_k having generated the orbit segment $X = \{x_i\}_{i=0}^N$. Note that $\alpha_j^k(n)$ can also be expressed

$$\alpha_j^k(n) = p(y(0), y(1), \cdots, y(n)|[x(n) \in I_j^k, f_k])\, p(x(n) \in I_j^k | f_k). \tag{7.6}$$

Also, we let $P^k = [p^k_{lj}]_{j,l=1}^{T_k}$ denote the TPM associated with f_k, where p^k_{lj} denotes the probability that $x(n) \in I^k_j$ given that $x(n-1) \in I^k_l$. (Note that P^k and p^k_{lj} have different meanings than in Chapter 6).

Using these definitions and notational conventions, we have the following computationally efficient algorithm, *the forward algorithm*, for calculating each likelihood function and optimally discriminating among the maps:

1. For each map f_k, compute $\alpha^k_j(n)$ as a function of n for $j = 1, \cdots, T_k$ with the following recursion:

$$\begin{aligned} \alpha^k_j(0) &= p(y(0)|[x(0) \in I^k_j, f_k])\, p(x(0) \in I^k_j | f_k) \\ \alpha^k_j(n+1) &= \left[\sum_{l=1}^{T_k} \alpha^k_l(n) p^k_{lj}\right] \\ &\quad \times p(y(n+1)|[x(n+1) \in I^k_j, f_k]), \quad n = 0, 1, \cdots, N-1 \end{aligned} \tag{7.7}$$

2. For each map, compute the likelihood $p(Y|f_k)$ by exploiting the relation given by

$$p(Y|f_k) = \sum_{j=1}^{T_k} \alpha^k_j(N). \tag{7.8}$$

3. Choose the map f_k for which $p(Y|f_k)$ is largest.

The detection algorithm requires specification of initial state probabilities $p(x(0) \in I^k_j | f_k)$ for each map. These probabilities are particularly easy to determine if the initial condition is a random variable with constant PDF over the unit interval. In this case, $p(x(0) \in I^k_j | f_k)$ is given by the length of I^k_j. In the more general case, $p(x(0) \in I^k_j | f_k)$ is given by the PDF of $x(0)$ integrated over I^k_j. The PDF of $x(0)$ can not be arbitrary, for the Markov chain property of the maps to hold as suggested by Corollary 1 in Chapter 6. In particular, the PDF of $x(0)$ must be constant over each partition element I^k_j for the Markov chain property to apply to map k. In general, when this condition is violated, the partition elements no longer correspond to the states of a homogeneous Markov chain, but they may correspond to the states of an inhomogeneous Markov chain, that is, a Markov chain with time-varying state transition probabilities.

7.2.3 Detection with Unquantized Orbit Points

When each of the M transformations h_k is the identity operator, the detection problem is that of discriminating among M EMC maps based on noisy observations of unquantized orbit points. In general, for the case of unquantized orbit points, optimal discrimination among the M maps is not computationally feasible because the initial condition $x(0)$ is unknown and the maps are chaotic. However, if the EMC maps each give rise to arbitrarily fine Markov partitions, computationally efficient, albeit suboptimal discrimination is still possible if we model the dynamics of each map as an HMM and apply the detection rule outlined earlier.

Specifically, we first select a sufficiently fine Markov partition for each map, with the the necessary fineness of the partition dependent upon the M maps being used. We know of no optimality criteria for quantifying this expression for a given set of maps. One suboptimal, practical method to choose partition sizes is by trial-and-error, with detection results obtained experimentally via Monte Carlo simulations. On the basis of experiments with various EMC maps which give rise to uniform Markov partitions, it appears that there is a threshold partition size, with finer partitions offering little if any improvement in performance over a partition with elements having length equal to this threshold size.

Having selected a Markov partition for each map, we apply the detection rule outlined in Section 7.2.2, with one fundamental change. In particular, we use the following expression for $o_j^k(n)$ in place of (7.2):

$$o_j^k(n) = u_j^k(n) + v(n), \tag{7.9}$$

where $u_j^k(n)$ is a random variable which is independent of $v(n)$ and which has a constant PDF with region of support I_j^k. Thus, each observation $y(n)$ is now the sum of two independent random variables with conditional PDF $p(y(n)|[x(n) \in I_j^k, f_k])$ given by the convolution of the PDFs of $u_j^k(n)$ and $v(n)$.

The motivation for modeling the noise-free output of each state as a uniform random variable with region of support over the corresponding partition element follows from two facts. First, as noted above, for the Markov chain property to hold, the PDF of the initial condition $x(0)$ must be constant over each partition element. Second, as shown in [12], the Frobenius-Perron operator restricted to PDFs that are constant over each partition element of a Markov partition for an MC map is a linear operator. As a consequence, if the PDF

of the initial condition $x(0)$ is constant over each partition element, the PDF induced by f acting on $x(0)$ is constant over each partition element as well. As a result, the PDF of $x(n)$ conditioned on $x(n)$ lying in a given partition element is constant over that partition element, thereby suggesting the appropriateness of modeling the noise-free output of each state of the Markov chain in the HMM as a uniform random variable.

It is straightforward to extend this suboptimal detection algorithm to the case in which each h_k is not the identity, but instead an arbitrary, piecewise differentiable, memoryless transformation. In this case, an appropriate model for the noise-free output of each state is that of a random variable with PDF induced by h_k acting on a uniformly distributed random variable, where the region of support of the uniformly distributed random variable is the corresponding partition element.

7.3 Detection Examples

In this section, we compare the performance of the optimal and suboptimal detection algorithms. We focus on the binary detection problem in which we seek to discriminate between two EMC maps based on a noise-corrupted orbit segment generated by one of the maps. For convenience but not by necessity, all performance results were obtained with a Gaussian, white noise sequence used for $\{v(i)\}_{i=0}^{N}$.

We first consider the two maps depicted in Figures 7-1 and 7-2, with typical orbit segments also shown in the figures. The maps are *antipodal* in the sense that for a given

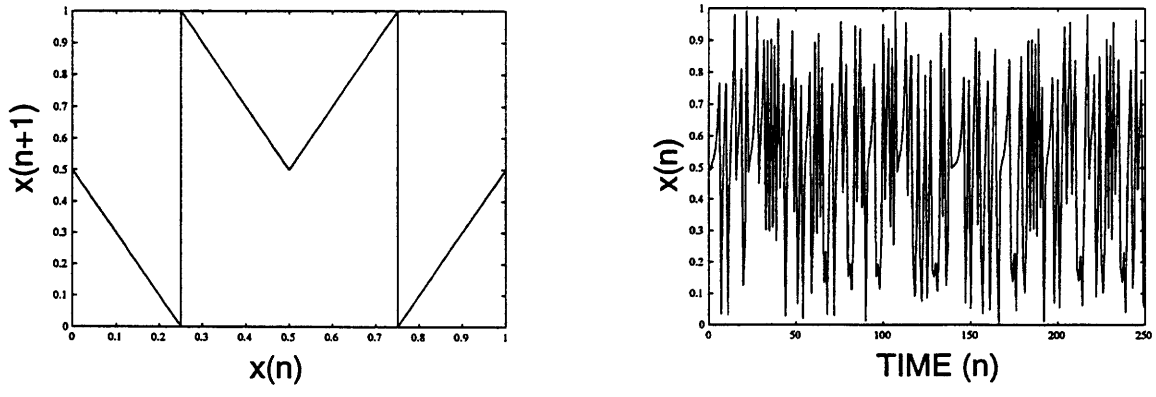

Figure 7-1: EMC map f_1 and typical orbit segment. (a) EMC map; (b) Orbit segment.

initial condition $x(0)$, corresponding orbit points of the maps satisfy the relation $f_1^i(x(0)) = 1 - f_2^i(x(0))$ for $i > 0$. (The conditions required for antipodality were established in [65]).

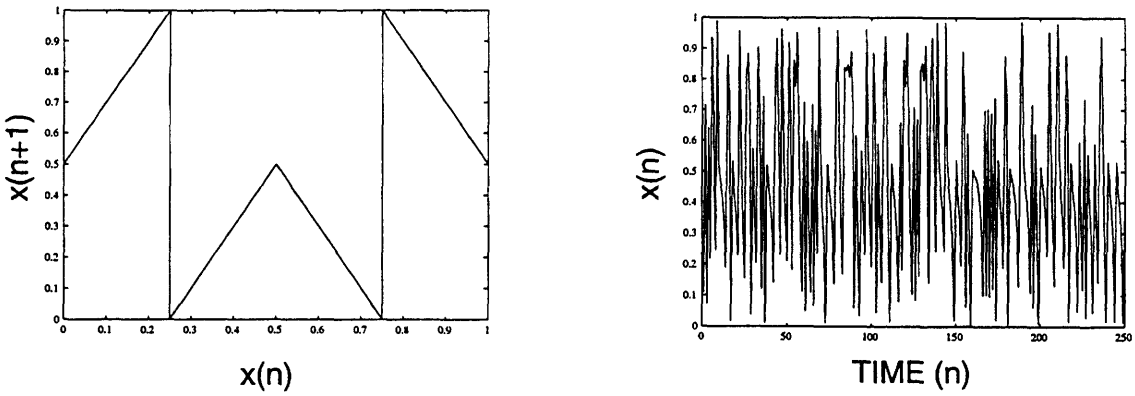

Figure 7-2: EMC map f_2 and typical orbit segment. (a) EMC map; (b) Orbit segment.

Because the maps are antipodal, their power density spectra are identical. Both maps also have identical stationary PDFs given by the constant value one over the unit interval. Any uniform partition of the unit interval into 4 N subintervals, where N is any positive integer, is a Markov partition for each map.

Figures 7-3 (a) and (b) depict the error probabilities as a function of the input SNR for an 8-element Markov partition and differently sized orbit segments.

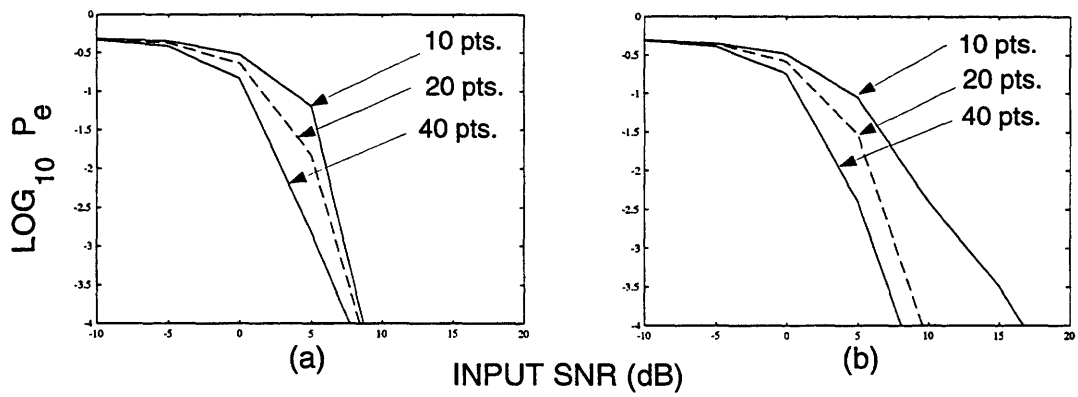

Figure 7-3: Detection error probabilities P_e for EMC maps f_1 and f_2 with 8-element Markov partitions and differently sized orbit segments used in the likelihood function. (a) Quantized orbit points; (b) Unquantized orbit points.

The actual ratio used for each input SNR value is that of the variance of a random variable with constant PDF over the unit interval (which equals 1/12) and the observation noise variance. Whereas the stationary PDF of each map has the constant value one over the unit interval, the SNR values are the same as the SNR values which use the actual signal variance (given by the variance of points in a typical orbit) when the orbits points are unquantized, but may differ slightly from the actual signal variance with coarsely quantized

orbit points. The curves are parameterized by the length of the noise-corrupted orbit segment used in the likelihood function. Ten thousand independent trials, with each map generating the orbit segments for half the trials, were used to obtain the error probabilities for values greater than or equal to .01, whereas one hundred thousand independent trials were used to obtain error probabilities below this value. The plotted results indicate that for a given input SNR, performance improves with an increase in the orbit segment size, as one might expect. In addition, the results suggest that the performance of the suboptimal detector (the one using unquantized orbit points) is comparable to that of the optimal detector (the one using quantized orbit points) with 20-point and 40-point orbit segments.

Figures 7-4 show performance results for differently sized Markov partitions and a 20-point orbit segment used in the likelihood function. The results suggest that the detection

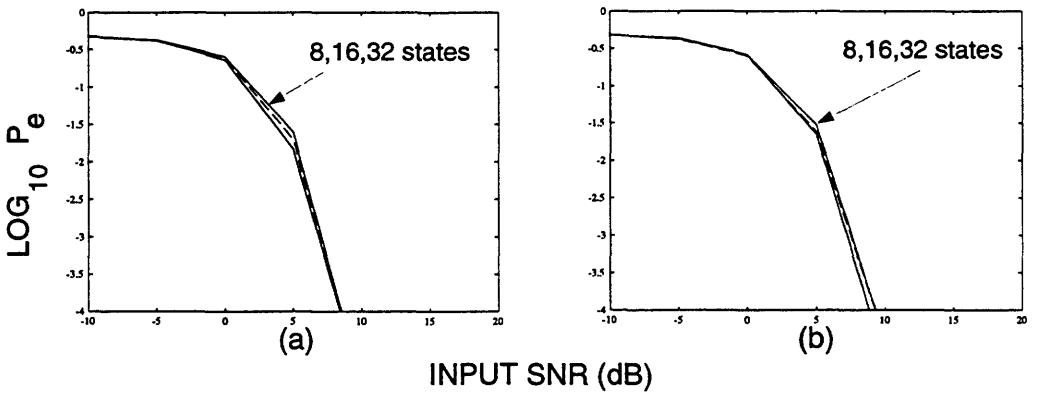

Figure 7-4: Detection error probabilities P_e for EMC maps f_1 and f_2 with 20-point orbit segments and differently sized Markov partitions used in likelihood functions. (a) Quantized orbit points; (b) Unquantized orbit points.

algorithm is insensitive to the size of the Markov partition used for discriminating between the two maps.

Figures 7-7 and 7-8 depict analogous performance results as those in Figures 7-3 and 7-4, but obtained with the two EMC maps f_3 and f_4 shown along with typical orbit segments in Figures 7-5 and 7-6. EMC maps f_3 and f_4 satisfy a weaker form of antipodality than maps f_1 and f_2. Specifically, while it is not true that $f_3^i(x(0)) = 1 - f_4^i(x(0))$ for $i > 0$, it is true that $f_3^i(x(0)) = 1 - f_4^i(1 - x(0))$. A comparison of the performance results for f_1 and f_2 with the results for f_3 and f_4 suggests that the former results are slightly better than the latter, as one might expect in light of the stronger form of antipodality satisfied by f_1 and f_2.

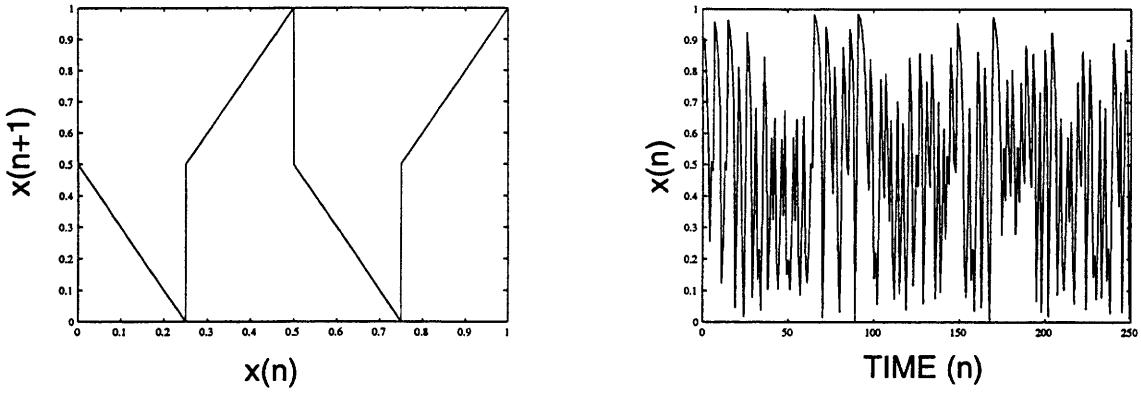

Figure 7-5: EMC map f_3 and typical orbit segment. (a) EMC map; (b) Orbit segment.

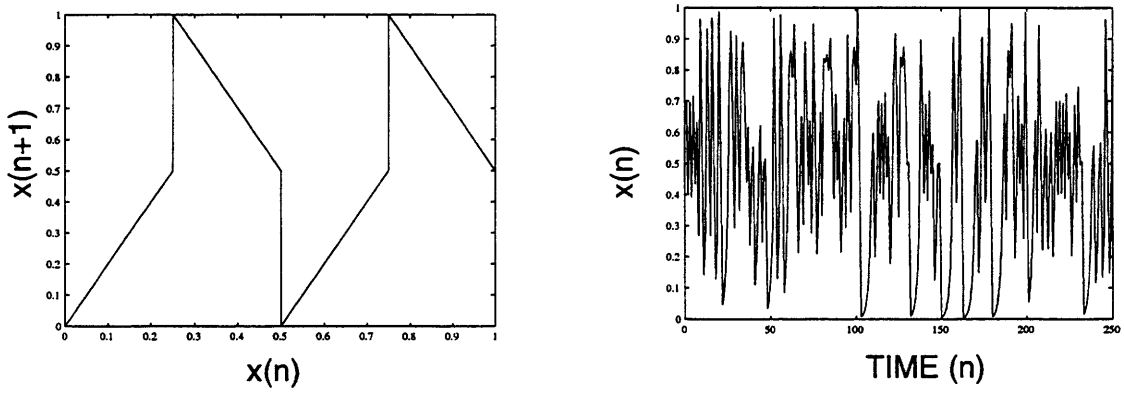

Figure 7-6: EMC map f_4 and typical orbit segment. (a) EMC map; (b) Orbit segment.

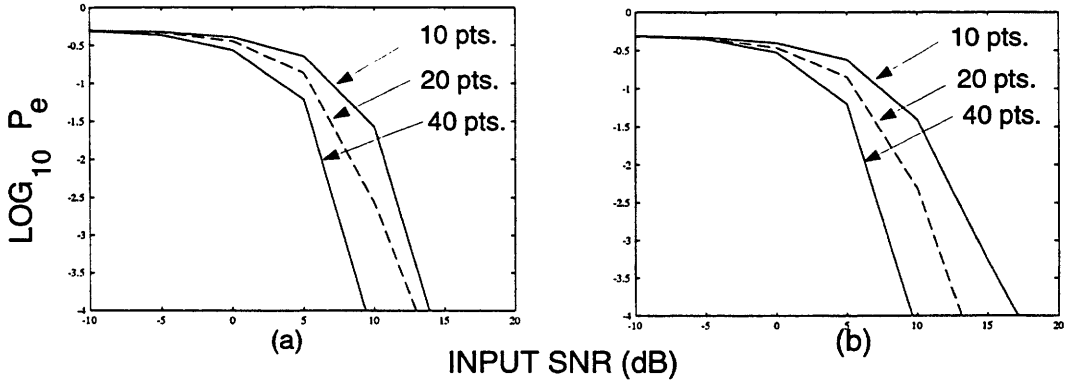

Figure 7-7: Detection error probabilities P_e for EMC maps f_3 and f_4 with 8-element Markov partitions and differently sized orbit segments used in likelihood functions. (a) Quantized orbit points; (b) Unquantized orbit points.

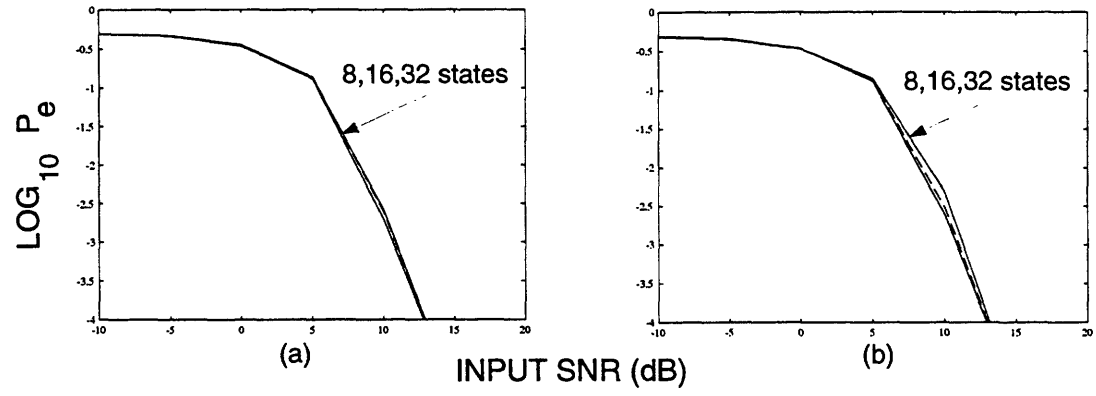

Figure 7-8: Detection error probabilities P_e for EMC maps f_3 and f_4 with 20-point orbit segments and differently sized Markov partitions used in likelihood functions. (a) Quantized orbit points; (b) Unquantized orbit points.

One should not infer from these examples that the detection algorithms are only useful with antipodal maps. In fact, there may be detection applications, such as secure communication over low-noise channels, in which the use of antipodal maps is neither necessary nor desirable.

7.4 Scale Factor and Noise Variance Estimation

The detection algorithms introduced in Sections 7.2.2 and 7.2.3 require that the observation noise variance as well as each output transformation h_k be known. However, these quantities may be unknown or only partially known in practical applications. For example, in applications in which orbit segments from EMC maps are transmitted over a channel, the orbit points may undergo a number of unknown transformations, such as linear and nonlinear filtering, fading, scaling, and noise corruption. In this section we focus on two of these transformations—scaling and noise corruption. In particular, we discuss how to simultaneously estimate a constant, multiplicative scaling factor applied to each orbit point and the variance of an additive, Gaussian, white, corrupting noise.

By exploiting the relation between noise-corrupted orbit segments of Markov maps and sample paths of hidden Markov models, we can derive computationally efficient iterative algorithms for estimating an unknown scale factor and noise variance. When the orbit points are properly quantized, the estimation algorithms are iterative ML estimation algorithms. With unquantized points, the algorithms are not ML estimation algorithms but are potentially effective, nevertheless. The specific estimation algorithms we use are variations of

the Baum-Welch re-estimation procedure [74, 75], an iterative ML technique for estimating parameters of hidden Markov models from sample observations. Of interest here is the case in which the EMC map is known, as well as a Markov partition for the map and the TPM of the Markov chain corresponding to the partition.

We first consider the case of quantized orbit points. For this case, we can formulate the re-estimation procedure for estimating the unknown noise variance and scale factor as follows. As noted above, we assume that we are given a Markov map f, a Markov partition for the map $\{I_j\}_{j=1}^T$, a TPM $\boldsymbol{P} = [p_{ij}]$ for the Markov chain associated with the partition, and an associated set of constant values $\{H_j\}_{j=1}^T$, with H_j the value associated with partition element I_j. The set $\{H_j\}_{j=1}^T$ consists of the the state values or equivalently the quantization values of the orbit points, with the value H_j associated with each orbit point lying in I_j. In the following equations, $H(n)$ denotes the quantization value associated with the orbit point $x(n)$, i.e., $H(n) = H_j$ if $x(n) \in I_j$. Finally, we assume that an $(N+1)$-point orbit segment $X = \{x(i)\}_{i=0}^N$ is generated by the map and we observe the observation set $Y = \{y(i)\}_{i=0}^N$, where $y(n)$ is given by

$$y(n) = k\,H(n) + v(n) \tag{7.10}$$

and where k is the unknown scale factor we seek to estimate and $\{v(i)\}_{i=0}^N$ is a Gaussian, white-noise sequence with unknown variance σ^2, which we also seek to estimate.

As in Sections 7.2.2 and 7.2.3, we define the forward variable $\alpha_j(n)$ as

$$\alpha_j(n) = p(y(0), y(1), \cdots, y(n), x(n) \in I_j), \tag{7.11}$$

which as indicated earlier is recursively computable:

$$\alpha_j(0) = p(y(0)|[x(0) \in I_j])\, p(x(0) \in I_j) \tag{7.12}$$

$$\alpha_j(n+1) = \left[\sum_{l=1}^{T} \alpha_l(n) p_{lj}\right] \times p(y(n+1)|[x(n+1) \in I_j]), \quad n = 0, 1, \cdots, N-1 \tag{7.13}$$

Now we also define the *backward variable* $\beta_j(n)$ as

$$\beta_j(n) = p(y(n+1), y(n+2), \cdots, y(N)|x(n) \in I_j), \tag{7.14}$$

which is recursively computable backward in time as follows:

$$\beta_j(N) = 1 \tag{7.15}$$

$$\beta_j(n) = \sum_{l=1}^{T} p_{jl}\, p(y(n+1)|[x(n+1) \in I_l])\, \beta_l(n+1), \quad n = 0, \cdots, N-1. \tag{7.16}$$

Finally, we define the *conditional state variable* $\gamma_j(n)$ as

$$\gamma_j(n) = p(x(n) \in I_j|Y), \tag{7.17}$$

where Y is the observation set $\{y(i)\}_{i=0}^{N}$. As indicated by (7.17), $\gamma_j(n)$ denotes the probability density that $x(n) \in I_j$ conditioned on the entire observation set. The forward, backward, and conditional state variables satisfy the following relation:

$$\gamma_j(n) = \frac{\alpha_j(n)\,\beta_j(n)}{p(Y)} = \frac{\alpha_j(n)\,\beta_j(n)}{\sum_{j=1}^{T} \alpha_j(n)\,\beta_j(n))} \tag{7.18}$$

where the denominator is simply a normalization constant.

It follows from more general results in [53, 75] that when each state S_j of a hidden Markov model has a scalar, Gaussian output PDF with mean m_j and variance σ_j^2, the re-estimation equations for $\hat{m}_j$ and $\hat{\sigma}_j^2$, the estimates of m_j and σ_j^2, are given by

$$\hat{m}_j = \frac{\sum_{i=0}^{N} \gamma_j(i)\, y(i)}{\sum_{i=0}^{N} \gamma_j(i)} \tag{7.19}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=0}^{N} \gamma_j(i)\, (y(i) - \hat{m}_j)^2}{\sum_{i=0}^{N} \gamma_j(i)} \tag{7.20}$$

Note that if $\gamma_j(i) = 1$ for all i, the re-estimation formulas are the sample mean and variance. The re-estimation formulas are iterative algorithms, with the estimated means and variances used to calculate the conditional state probabilities $\{\gamma_j(n)\}$ which are subsequently used in the above equations to re-estimate the means and variances.

For the problem of interest here, $m_j = k\, H_j$ where k is the unknown scale factor and H_j is known. For this problem, a straightforward derivation, (for which we know of no references) analogous to that used in [53] to obtain the above re-estimation equations,

yields the following re-estimation equation for $\hat{k}$, the estimate of the scale factor k:

$$\hat{k} = \frac{\sum_{i=0}^{N} \sum_{j=1}^{T} \gamma_j(i)\, H_j\, y(i)}{\sum_{i=0}^{N} \sum_{j=1}^{T} \gamma_j(i)\, H_j^2} \tag{7.21}$$

which is a weighted average of the observations, as is the case with (7.19), but with the weights now also dependent on the known values $\{H_j\}$. Similarly, since $\sigma_j^2 = \sigma^2$ is the same for each state, a straightforward derivation yields the following re-estimation equation for $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{\sum_{i=0}^{N} \sum_{j=1}^{T} \gamma_j(i)\, (y(i) - \hat{k}\, H_j)^2}{\sum_{i=0}^{N} \sum_{j=1}^{T} \gamma_j(i)} \tag{7.22}$$

$$= \frac{1}{N+1} \sum_{i=0}^{N} \sum_{j=1}^{T} \gamma_j(i)\, (y(i) - \hat{k}\, H_j)^2. \tag{7.23}$$

As discussed in [75], the Baum-Welch re-estimation procedure is closely related to the Expectation-Maximization (EM) algorithm, and as with the EM algorithm only local convergence is assured with the re-estimation procedure. As a consequence, the estimated parameter values converge to steady-state values which may or may not be the actual parameter values.

Two problems arise with unquantized orbit points. First, the HMM model introduced in Section 7.2.3 is not an exact representation of the dynamics underlying the observations. Second, the derivation of closed form re-estimation equations for an unknown scale factor and observation noise variance appears to be an intractable problem. However, using the above re-estimation equations for quantized orbit points as a foundation, we can derive Baum-Welch-like re-estimation equations for the unknown scale factor and noise variance with the HMM model for unquantized orbit points. Recall that with this model, each observation $y(n)$ is the sum of two random variables, one the observation noise term with unknown variance we seek to estimate, and the other a uniform random variable with mean $k\, L_j + \frac{k\, L_{j+1} - k\, L_j}{2}$ and variance $\frac{(k\, L_{j+1} - k\, L_j)^2}{12}$ for some j, where L_j denotes the left endpoint of partition element I_j and k denotes the unknown scale factor. Thus, an intuitively reasonable re-estimation procedure for obtaining the unknown scale factor and variance is to use $k\, L_j + \frac{k\, L_{j+1} - k\, L_j}{2}$ in place of $k\, H_j$ in the above re-estimation equations for quantized orbit points and to subtract from each term in the above expression for the variance estimate, an estimate of the variance of the uniform random variable associated with that term, or in other words

a term having the form $\frac{(k\,L_{j+1}-k\,L_j)^2}{12}$. The "Baum-Welch-like" re-estimation equations for unquantized orbit points that result are the following:

$$\hat{k} = \frac{\sum_{i=0}^{N}\sum_{j=1}^{T}\gamma_j(i)\,(L_j+\frac{L_{j+1}-L_j}{2})\,y(i)}{\sum_{i=0}^{N}\sum_{j=1}^{T}\gamma_j(i)\,(L_j+\frac{L_{j+1}-L_j}{2})^2} \tag{7.24}$$

$$\hat{\sigma}^2 = \frac{1}{N+1}\sum_{i=0}^{N}\sum_{j=1}^{T}\gamma_j(i)\times \tag{7.25}$$

$$\left[\left(y(i)-\hat{k}\,(L_j+\frac{L_{j+1}-L_j}{2})\right)^2-\frac{(\hat{k}\,L_{j+1}-\hat{k}\,L_j)^2}{12}\right] \tag{7.26}$$

where we define L_{N+1} to be the right endpoint of I_N.

Figures 7-9 and 7-10 depict the error in estimating the scale factor and noise variance for the EMC map shown in Figure 7-5. The actual scale factor used for the results was 2 and the

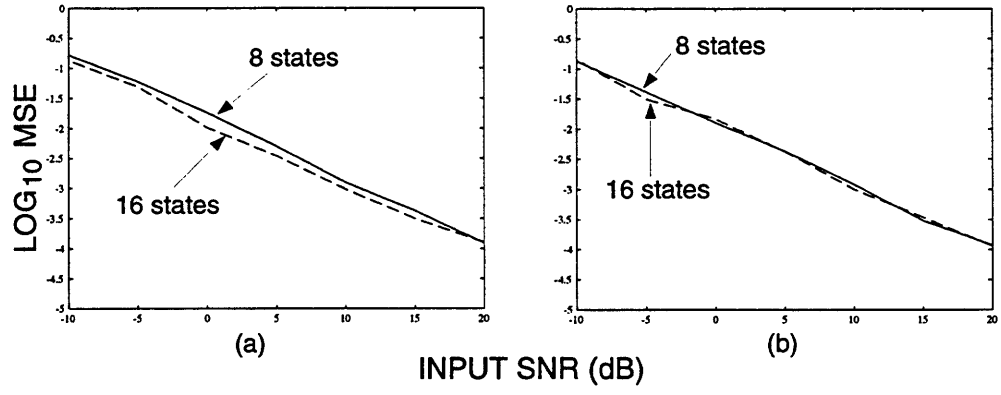

Figure 7-9: Mean-squared error (MSE) (average of 100 trials) for scale factor estimation for EMC f_3 map using 100 observations. (Actual value of scale factor value was 2). (a) Quantized orbit points; (b) Unquantized orbit points.

noise variance was determined by the SNR. The performance measure used in Figures 7-9 (a) and (b) is the squared estimation error averaged over 100 independent trials. Similarly, the performance measure used in Figures 7-10 (a) and (b) is the squared estimation error averaged over 100 independent trials and normalized by the actual value of the variance. The curves in the figures are parameterized by the size of the uniform Markov partition used, with 100 observations used in the estimation equations for each plotted result. The scale factor and variance estimates were both initialized with the value 1. The estimation algorithms were iterated either until they converged or an upper limit of 50 iterations was reached.

As indicated by the figures, the results with quantized and unquantized orbit points are

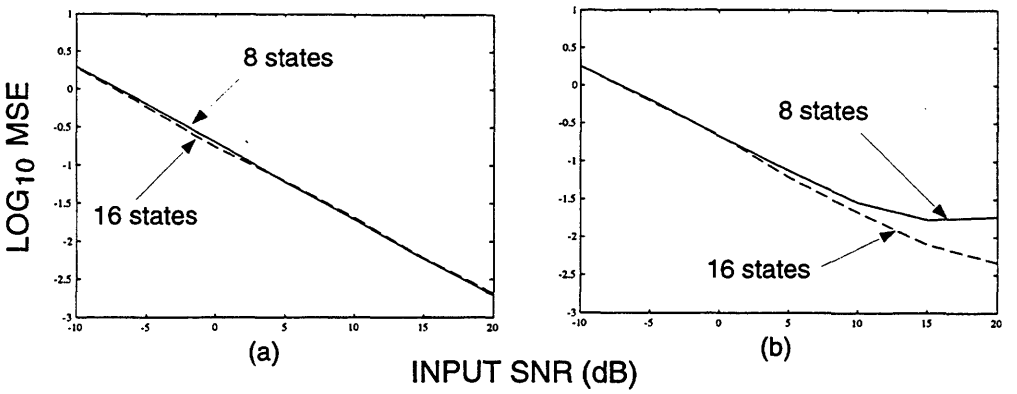

Figure 7-10: Normalized mean-squared error (MSE) (average of 100 trials) for noise variance estimation for EMC map f_3 using 100 observations. (Noise variance determined by input SNR). (a) Quantized orbit points; (b) Unquantized orbit points.

comparable except for variance estimation with large input SNRs and an 8-element Markov partition, or equivalently an 8-state Markov chain in the HMMs. The results also indicate an insensitivity to the Markov partition size, except for variance estimation at high input SNRs with unquantized orbit points. Figures 7-11 and 7-12 depict the performance results for a fixed-sized, 16-element uniform Markov partition (equivalently a 16-state Markov chain in the HMMs) and differently sized orbit segments. The figures suggest a strong correlation between scale factor estimation accuracy and the orbit segment size but negligible correlation between variance estimation accuracy and the orbit segment size.

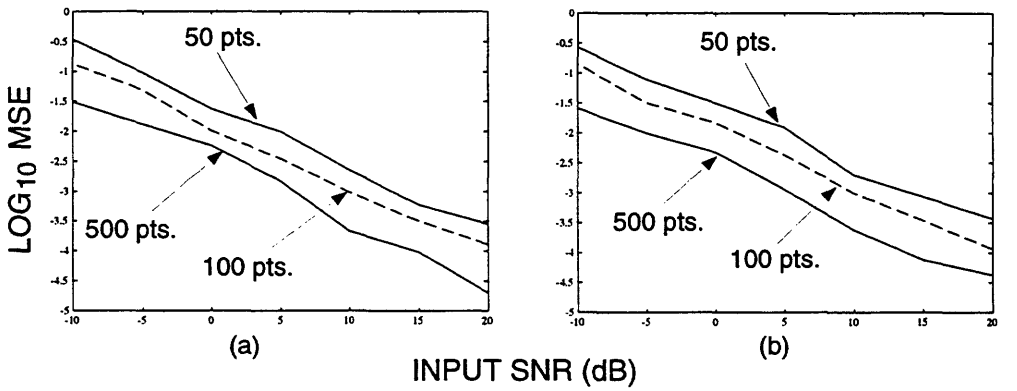

Figure 7-11: Mean-squared error (MSE) for scale factor estimation for EMC map f_3 using 16-element uniform Markov partition. (Actual value of scale factor value was 2). (a) Quantized orbit points; (b) Unquantized orbit points.

The question arises as to the advantage of the re-estimation equations for unquantized orbit points over the re-estimation equations for quantized orbit points, when the observations arise from unquantized orbit points. Figure 7-13 depicts the absolute bias, defined

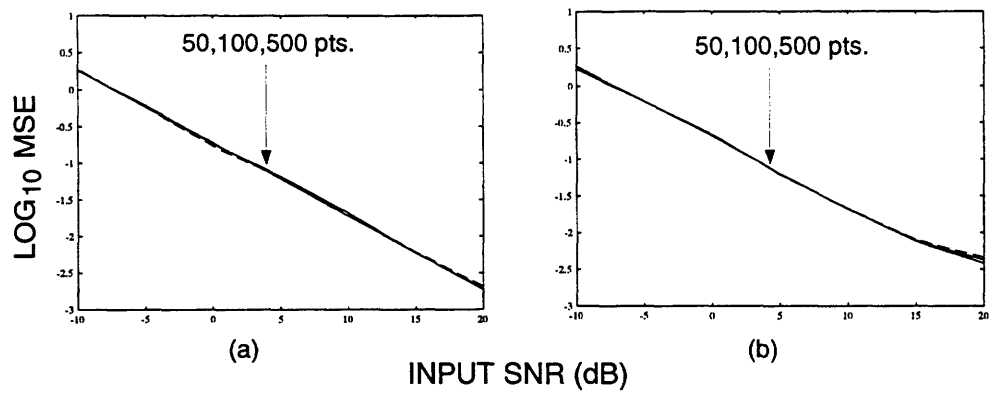

Figure 7-12: Normalized mean-squared error (MSE) for noise variance estimation for EMC map f_3 using 16-element uniform Markov partition. (Noise variance determined by input SNR). (a) Quantized orbit points; (b) Unquantized orbit points.

as the absolute value of the difference between the estimated scale factor and actual scale factor, averaged over 100 independent trials, obtained with both scale factor re-estimation equations when applied to observations arising with unquantized orbit points. As indicated

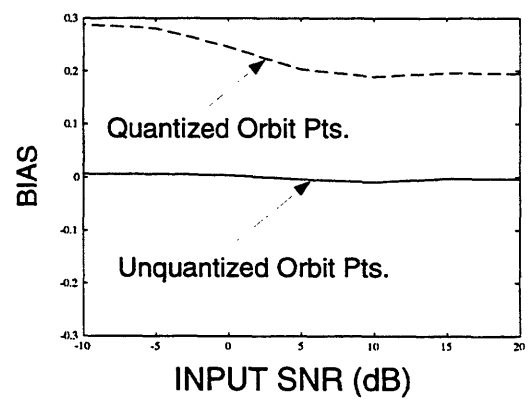

Figure 7-13: Estimated scale factor minus actual scale factor with re-estimation equations for quantized and unquantized orbit points using 100 observations arising from unquantized orbit points and a 16-element uniform Markov partition. (Actual value of scale factor was 2).

in the figure, a nonnegligible positive bias arises at all SNRs with the re-estimation equation for quantized orbit points. Additional experiments have shown the bias to be dependent on the scale factor and to increase as the scale factor increases.

7.5 Map Discriminability

A useful tool when selecting EMC maps for detection applications would be a metric which quantifies the discriminability among these maps. Unfortunately, finding such a metric remains an elusive goal. We have explored metrics based on divergence (or cross entropy) rates of Markov chains [19] and Bhattacharya distance measures [41] with little success. The fundamental problem is in finding a metric that is consistent with experimentally obtained detection results.

A useful empirical metric for hidden Markov models with finite observation alphabets was introduced in [39]. This metric, which uses asymptotic cross-information rates (defined below), is easily adapted for use with continuous-valued observations and EMC maps. Undesirable aspects of the metric are that it is a function of the observation noise variance and its calculation may require that a large set of simulated observations be obtained for each map being compared.

The metric arises as follows. We start with a set of M hidden Markov models $\{f_i\}$ with each consisting of an underlying homogeneous, finite-state Markov chain and state-dependent output. For the problems of interest here, the underlying Markov chains are those corresponding to Markov partitions of EMC maps and the output associated with each state of the Markov chain is either a Gaussian random variable when orbit points are quantized or the sum of a Gaussian random variable and uniformly distributed random variable when orbit points are not quantized.

We let Y_n^j denote a set of n observations associated with f_j and define an n-point information rate $E(j,j,n)$ as

$$E(j,j,n) \equiv \frac{1}{n} \log p(Y_n^j | f_j) \tag{7.27}$$

and n-point cross-information rates $E(k,j,n)$ as

$$E(k,j,n) \equiv \frac{1}{n} \log p(Y_n^j | f_k), \quad k = 1,2,\cdots,M,\ k \neq j. \tag{7.28}$$

These rates are simply log-likelihood values normalized by the number of observations. Note that these quantities are not entropy rates as they do not involve the expectation over the observation set.

For a restricted form of this scenario in which only a finite number of values are possible for each output, it was shown in [70] that each of the following limits exists:

$$E(k,j) \equiv \lim_{n\to\infty} \frac{1}{n} \log p(Y_n^j | f_k), \quad k = 1, \cdots M \tag{7.29}$$

and that the following relations hold:

$$E(j,j) \geq E(k,j), \quad k = 1, \cdots M. \tag{7.30}$$

For the detection problems of interest here involving continuous-valued outputs, convergence of the n-point information rates has never been proven. Nonetheless, experimental results obtained with EMC maps suggest that the n-point information and cross-information rates approach asymptotic values with small perturbations about these values as n grows large. For example, Figure 7-14 depicts differences of information and cross-information rates for EMC maps f_1 and f_2 shown in Figures 7-1 and 7-2 as a function of time with an input SNR of 0 dB. As the figures suggest, the differences appear to approach asymptotic mean

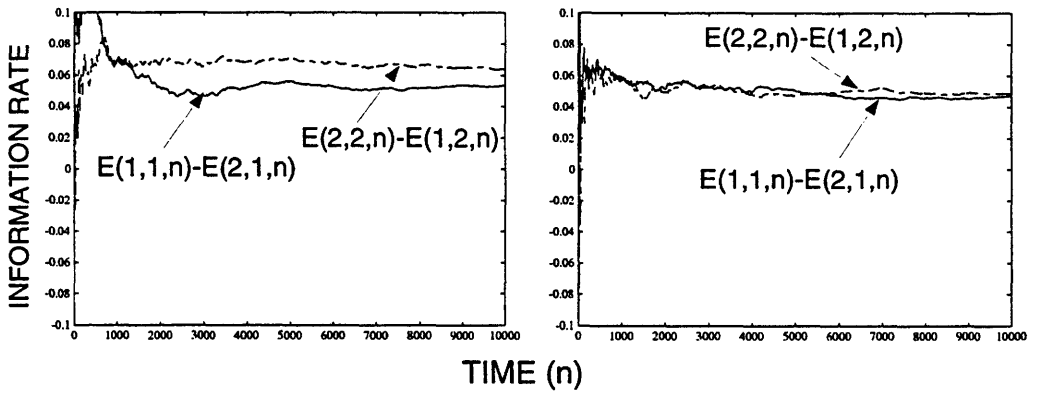

Figure 7-14: Information rate differences as a function of time for EMC maps f_1 and f_2 with an input SNR of 0 dB. (a) Quantized orbit points; (b) Unquantized orbit points.

values with small perturbations about the mean. The motivation for showing the rate differences $E(1,1,n) - E(2,1,n)$ and $E(2,2,n) - E(1,2,n)$ and not the individual information and cross-information rates arises from the close relation among these differences and the discrimination criterion used in the detection algorithms introduced earlier. In particular, using (7.27) and (7.28), we have the following:

$$E(1,1,n) - E(2,1,n) \;=\; \log \frac{p(Y_n^1 | f_1)}{p(Y_n^1 | f_2)} \tag{7.31}$$

$$E(2,2,n) - E(1,2,n) = \log \frac{p(Y_n^2|f_2)}{p(Y_n^2|f_1)} \tag{7.32}$$

With the detection algorithms, the discrimination criterion consists of deciding upon that map for which the corresponding likelihood $p(Y_n^j|f_k)$ is largest for a fixed observation set Y_n^j. For binary detection involving maps f_1 and f_2, a correct decision is made when (7.31) is positive for the case in which f_1 generated the observations. Similarly, a correct decision is made when (7.32) is positive for the case in which f_2 generated the observations. As such, the differences $E(1,1,n) - E(2,1,n)$ and $E(2,2,n) - E(1,2,n)$ are qualitative indicators of the discriminability between the maps as n grows large, with larger, positive values of the differences suggesting greater discriminability.

Figure 7-15 depicts information rate differences for EMC maps f_3 and f_4 shown in Figures 7-5 and 7-6 as a function of time with an input SNR of 0 dB. As the figures suggest

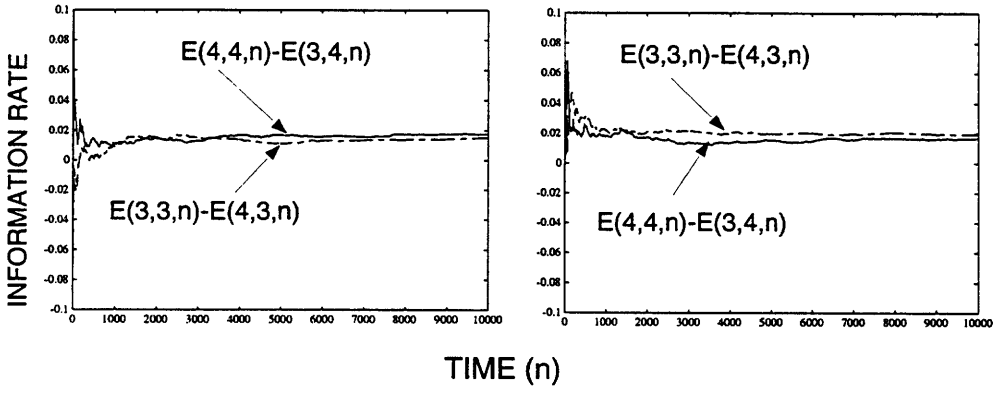

Figure 7-15: Information rate differences as a function of time for EMC maps f_3 and f_4 with an input SNR of 0 dB. (a) Quantized orbit points; (b) Unquantized orbit points.

and as the performance results presented earlier confirm, the discriminability between f_1 and f_2 is much greater than that between f_3 and f_4

7.6 State Estimation with MC Maps

7.6.1 Problem Overview

In the preceding sections of this chapter, we have shown how the close relation between EMC maps and Markov chains facilitates the detection or discrimination of these maps based on noise-corrupted orbit segments. In this section, we show how this relation facilitates practical, optimal and suboptimal ML state estimation with EMC maps and more

generally MC maps.

We focus on the fixed-interval smoothing scenario in which we are given an MC map f, an unobserved $(N+1)$-point orbit segment $\{x(i) = f^i(x(0))\}_{i=0}^N$ generated by f, and a set of $N+1$ observations $Y = \{y(i)\}_{i=0}^N$ with $y(i)$ given by

$$y(i) = x(i) + v(i) \tag{7.33}$$

where $\{v(i)\}_{i=0}^N$ is a Gaussian, white-noise sequence with variance σ^2. It follows from earlier results in the thesis that for this scenario, the log-likelihood function for the j^{th} orbit point, $\log p(Y; x(j))$, is given by

$$\log p(Y; x(j)) = C(N) - \sum_{i=0}^{N} \frac{(y(i) - f_{(x(0),j)}^{i-j}(x(j)))^2}{2\sigma^2} \tag{7.34}$$

where $C(N)$ is a constant independent of both the observation set Y and the orbit segment, and where $f_{(x(0),j)}^{i-j}(x(j))$ equals $f^{i-j}(x(j))$ if $i - j \geq 0$ and equals the actual value of the inverse image $f^{i-j}(x(j))$ which gave rise to the observations if $i - j < 0$. In other words, we assume that there is a fixed initial condition $x(0)$ so that the following holds:

$$x(i) = f^i(x(0)) = f_{(x(0),j)}^{i-j}(x(j)) \quad i, j = 0, \cdots, N \tag{7.35}$$

Because f is a deterministic mapping, there is a bijective correspondence between initial conditions $x(0)$ and $(N+1)$-point orbit segments $\{f^i(x(0))\}_{i=0}^N$, so that any property associated with a specific orbit segment can be associated with a specific initial condition as well.

We let L denote the *smallest* number of affine segments of f in the sense that two affine segments are considered part of the same segment if they have the same affine parameter pair. Also, we let (τ_i, β_i) and A_i denote the affine parameter pair and segment domain, respectively, associated with the i^{th} affine segment, so that $f(x) = \tau_i x + \beta_i$ if $x \in A_i$. Note that each A_i is a subinterval of the unit interval, disjoint from all the other A_i, and the set of subintervals $\{A_i\}_{i=1}^L$ is a partition of the unit interval but not necessarily a Markov partition of f.

We can associate sequences of $N+1$ affine parameter pairs and affine segment domains with each $(N+1)$-point orbit segment or equivalently with each initial condition. In par-

ticular, for the initial condition x, we associate the sequences $\{(\tau(i,x),\beta(i,x))\}_{i=0}^{N}$ and $\{A(i,x)\}_{i=0}^{N}$ where $\tau(i,x) = \tau_j$, $\beta(i,x) = \beta_j$, and $A(i,x) = A_j$ if $f^i(x) \in A_j$. One can show that because each A_i is a subinterval and disjoint from all the other A_i, the set of initial conditions with the same associated sequence of segment domains $\{A(i,x)\}_{i=0}^{N}$ is a subinterval $A(x)$, which is given by

$$A(x) = \bigcap_{i=0}^{N} f^{-i}(A(i,x)) \tag{7.36}$$

where f^{-i} denotes the possibly multiple-valued, inverse image of the composed map f^i.

It is straightforward but tedious to show that because f is piecewise linear, f^i is piecewise linear as well, so that

$$f^i(x) = T(i,0,x)\,x + B(i,0,x) \quad i = 0,\cdots,N \tag{7.37}$$

where

$$T(0,0,x) = 1 \tag{7.38}$$

$$B(0,0,x) = 0 \tag{7.39}$$

and for $i > 0$

$$T(i,0,x) = \prod_{k=0}^{i-1} \tau(k,x) \tag{7.40}$$

$$B(i,0,x) = \beta(i-1,x) + \sum_{k=0}^{i-2} \left[\beta(k,x) \prod_{l=k+1}^{i-1} \tau(l,x) \right]. \tag{7.41}$$

In addition, for a given orbit segment $\{x(i) = f^i(x)\}_{i=0}^{N}$, the following holds

$$f^{i-j}_{(x(0),j)}(x(j)) = T(i,j,x)\,x(j) + B(i,j,x) \quad i,j = 0\cdots,N \tag{7.42}$$

where

$$T(j,j,x) = 1 \tag{7.43}$$

$$B(j,j,x) = 0 \tag{7.44}$$

while for $i > j$

$$T(i,j,x) = \prod_{k=0}^{i-j-1} [\tau(k+j,x)] \tag{7.45}$$

$$B(i,j,x) = \beta(i-1,x) + \sum_{k=0}^{i-j-2} \left[\beta(k+j,x) \prod_{l=k+1}^{i-j-1} \tau(l+j,x) \right]. \tag{7.46}$$

and for $i < j$

$$T(i,j,x) = \frac{1}{\prod_{k=0}^{j-i-1} \tau(i+k,x)} \tag{7.47}$$

$$B(i,j,x) = -T(i,j,x) \times \left\{ \beta(j-1,x) + \sum_{k=0}^{j-i-2} \left[\beta(i+k,x) \prod_{l=k+1}^{j-i-1} \tau(i+l,x) \right] \right\} \tag{7.48}$$

Therefore, we can express (7.34) as

$$\log p(Y; x(j)) = C(N) - \sum_{i=0}^{N} \frac{(y(i) - T(i,j,x)\, x(j) - B(i,j,x))^2}{2\sigma^2}, \tag{7.49}$$

which is a quadratic function of the unknown orbit point $x(j)$.

7.6.2 ML State Estimation Considerations

We first consider maximizing (7.49) for the special case in which $j = 0$, so that $x(j) = x(0) \equiv x$ and (7.49) becomes

$$\log p(Y; x) \equiv \log p(Y; x(0)) = C - \sum_{i=0}^{N} \frac{(y(i) - T(i,0,x)\, x - B(i,0,x))^2}{2\sigma^2}, \tag{7.50}$$

If the sequence of parameter pairs $\{(T(i,0,x), B(i,0,x))\}_{i=0}^{N}$ is known and independent of x, then finding the value(s) of x for which (7.50) has extremal values for a given observation set is a straightforward calculus problem having a closed-form solution. In this special case, (7.50) has a unique extremum (except possibly for degenerate cases) because it is quadratic in x. Furthermore, this extremum is a maximum because (7.49) becomes arbitrarily small as x becomes arbitrarily large or small.

Now consider the case in which the affine parameter pairs $\{(T(i,0,x), B(i,0,x))\}_{i=0}^{N}$ are those associated with $\hat{x}_{ML}$, the ML estimate of the initial condition x. In other words, as with any initial condition, associated with $\hat{x}_{ML}$ is a an orbit segment $\{f^i(\hat{x}_{ML})\}_{i=0}^{N}$, a

sequence of affine parameter pairs $\{(\tau(i,\hat{x}_{ML}),\beta(i,\hat{x}_{ML}))\}_{i=0}^{N}$, and a sequence of affine segment domains $\{A(i,\hat{x}_{ML})\}_{i=0}^{N}$. In addition, there is a sequence of parameter pairs $\{(T(i,0,\hat{x}_{ML}),B(i,0,\hat{x}_{ML}))\}_{i=0}^{N}$ uniquely determined from $\{(\tau(i,\hat{x}_{ML}),\beta(i,\hat{x}_{ML}))\}_{i=0}^{N}$. Substituting the parameter pairs $\{(T(i,0,\hat{x}_{ML}),B(i,0,\hat{x}_{ML}))\}_{i=0}^{N}$ in (7.50) and maximizing over x yields the following expression for the maximizing value x_{MAX}:

$$x_{MAX}=\frac{\sum_{i=0}^{N}T(i,0,\hat{x}_{ML})(y(i)-B(i,0,\hat{x}_{ML}))}{\sum_{i=0}^{N}T^{2}(i,0,\hat{x}_{ML})}. \tag{7.51}$$

We now show that x_{MAX} equals $\hat{x}_{ML}$ if it is in $A(\hat{x}_{ML})$, the subinterval of initial conditions with the associated sequence of affine segment domains $\{A(i,\hat{x}_{ML})\}_{i=0}^{N}$. If not, $\hat{x}_{ML}$ is the endpoint of $A(\hat{x}_{ML})$ for which (7.50) has the larger value. The validity of this assertion follows from the fact that by definition x_{MAX} is the value of x that maximizes (7.50) for the fixed sequence of parameter pairs $\{(T(i,0,\hat{x}_{ML}),B(i,0,\hat{x}_{ML}))\}_{i=0}^{N}$. Also by definition, $\hat{x}_{ML}$ is the value of x that maximizes (7.50) when the expression is evaluated with the correct parameter pairs $\{(T(i,0,x),B(i,0,x))\}_{i=0}^{N}$ associated with x. Since (7.50) has a single maximum for a given sequence of parameter pairs and $\{(T(i,0,\hat{x}_{ML}),B(i,0,\hat{x}_{ML}))\}_{i=0}^{N}$ is the sequence associated with $\hat{x}_{ML}$, it follows that x_{MAX} equals $\hat{x}_{ML}$ when the former is in the subinterval of initial conditions associated with $\{(T(i,0,\hat{x}_{ML}),B(i,0,\hat{x}_{ML}))\}_{i=0}^{N}$. However, it is possible that x_{MAX} lies outside this subinterval. In this case, because (7.50) is a quadratic function of x (for fixed sequence of parameter pairs) and because $\hat{x}_{ML}\in A(\hat{x}_{ML})$, it follows that $\hat{x}_{ML}$ is the endpoint of $A(\hat{x}_{ML})$ closer to x_{MAX} or equivalently the endpoint for which (7.50) has the larger value.

What we have shown thus far is that if $\{(\tau(i,\hat{x}_{ML}),\beta(i,\hat{x}_{ML}))\}_{i=0}^{N}$, the sequence of affine parameter pairs associated with the ML orbit segment, is known one can in theory determine the ML orbit segment by first evaluating 7.51 to obtain x_{MAX}. If $x_{MAX}\in A(\hat{x}_{ML})$, then $\hat{x}_{ML}$, the ML estimate of $x(0)$, equals x_{MAX}; if not, then $\hat{x}_{ML}$ equals the endpoint of $A(\hat{x}_{ML})$ closer to x_{MAX}. Having obtained $\hat{x}_{ML}$, one can obtain $x_{ML}(i)$, the ML estimate of the i^{th} point of the orbit segment, by using the relation $x_{ML}(i)=f^{i}(\hat{x}_{ML})$. However, in practice it is often difficult to determine $A(\hat{x}_{ML})$. An additional practical concern is that the theoretical relation $x_{ML}(i)=f^{i}(\hat{x}_{ML})$ is generally not a useful relation with EMC maps, since these maps are expanding and thus inevitable computer round-off error in the determination of $\hat{x}_{ML}$ is amplified by successive compositions of the maps with themselves.

We now consider an alternative approach for calculating the ML orbit segment. The approach is equivalent in theory to the approach just discussed, but it is more useful in practice. It involves first obtaining $\hat{x}_{ML}(N)$, the ML estimate of the final point of the orbit segment, instead of $\hat{x}_{ML}$, the ML estimate of the first point of the orbit segment. The approach can be used with any EMC map for which the image of each affine segment domain under f is a union of affine segment domains and the slope of each affine segment is an integer with absolute value greater than one.

Because f is deterministic and x_{MAX} is the value of $x(0)$ that maximizes (7.49) for $j = 0$ when $T(i,0,x) = T(i,0,\hat{x}_{ML})$, and $B(i,0,x) = B(i,0,\hat{x}_{ML})$ for each $i = 0,\cdots,N$, it follows that $f^N(x_{MAX})$ is the value of $x(N)$ that maximizes (7.49) for $j = N$ when $T(i,N,x) = T(i,N,\hat{x}_{ML})$ and $B(i,N,x) = B(i,N,\hat{x}_{ML})$ for each i. However, direct maximization of (7.49) for the fixed sequence of affine parameter pairs $\{T(i,N,\hat{x}_{ML}), B(i,N,\hat{x}_{ML})\}$ yields the following for the maximizing value $x_{MAX}(N)$:

$$x_{MAX}(N) = \frac{\sum_{i=0}^{N} T(i,N,\hat{x}_{ML})(y(i) - B(i,N,\hat{x}_{ML}))}{\sum_{i=0}^{N} T^2(i,N,\hat{x}_{ML})}. \tag{7.52}$$

Since $f^N(x_{MAX})$ and $x_{MAX}(N)$ both maximize $\log p(Y; x(N))$ which is a quadratic expression in $x(N)$ with a single extremum, it follows that $x_{MAX}(N) = f^N(x_{MAX})$ and thus $x_{MAX} = f^{-N}_{(\hat{x}_{ML},N)}(x_{MAX}(N))$ where $f^N_{(\hat{x}_{ML},N)}$ denotes the invertible mapping implicitly defined by the sequence of affine parameter pairs $\{\tau(i,\hat{x}_{ML}), \beta(i,\hat{x}_{ML})\}_{i=0}^{N}$. Thus, one can determine x_{MAX} by first evaluating (7.52) and then using the relation $x_{MAX} = f^{-N}_{(\hat{x}_{ML},N)}(x_{MAX}(N))$.

The question arises as to the relevance and value of this alternative method for calculating x_{MAX}. There is a twofold answer. First, given any MC map f, not necessarily an EMC map, for which the image of each affine segment domain under f is a union of affine segment domains and given any two segment domains A_i and A_j satisfying $A_i \cap f^{-1}(A_j) \neq \emptyset$, then $f^{-1}(x) \cap A_i \neq \emptyset$ for each $x \in A_j$. In other words, since $f(A_i)$ is a union of affine segment domains, it follows that if A_j is in that union of segment domains and $x \in A_j$, then at least one point in the possibly multiple point set $f^{-1}(x)$ lies in A_i.

It further follows by induction that if f has this property of affine segment domains mapping onto unions of affine segment domains and $y \in A(N,\hat{x}_{ML})$, where $A(N,\hat{x}_{ML})$ is the affine segment domain containing $f^N(\hat{x}_{ML})$, then $f^{-N}(y) \cap A(0,\hat{x}_{ML}) \neq \emptyset$ and more

importantly $f^{-N}(y) \cap A(\hat{x}_{ML}) \neq \emptyset$ as well as $f^{-N}_{(\hat{x}_{ML},N)}(y) \in A(\hat{x}_{ML})$. What this means is that an equivalent condition for checking if $x_{MAX} \in A(\hat{x}_{ML})$ is to check if $x_{MAX}(N) \in A(N, \hat{x}_{ML})$, a far simpler task since $A(N, \hat{x}_{ML})$ is an affine segment domain and these domains are known *a priori*.

As a result, if $x_{MAX}(N) \in A(N, \hat{x}_{ML})$, then $x_{MAX} = f^{-N}_{(\hat{x}_{ML},N)}(x_{MAX}(N)) \in A(\hat{x}_{ML})$ and thus $\hat{x}_{ML} = x_{MAX} = f^{-N}_{(\hat{x}_{ML},N)}(x_{MAX}(N))$. If $x_{MAX}(N) \notin A(N, \hat{x}_{ML})$, then by a similar argument as used earlier $\hat{x}_{ML}(N) \equiv f^N(\hat{x}_{ML})$ is given by the endpoint z of $A(N, \hat{x}_{ML})$ for which $\log p(Y; x(N))$ has the larger value and $\hat{x}_{ML} = f^{-N}_{(\hat{x}_{ML},N)}(z)$.

The second part of the twofold answer to the question as to the value of using $x_{MAX}(N)$ to determine x_{MAX} involves the constraint that f be an EMC map for which the slope of each affine segment is an integer with absolute value greater than one. As discussed in Section 6.7, such an EMC map has a recoverability or invertibility property in the sense that the entire orbit segment is recoverable from the final orbit point in both a practical and theoretical sense. Therefore, given such an EMC map, we can recover $\hat{x}_{ML}$ and the entire ML orbit segment $\{f^i(\hat{x}_{ML})\}_{i=0}^N$ from $\hat{x}_{ML}(N)$ in theory and practice if the *inverse* system $\{f^{-i}_{(\hat{x}_{ML},N)}\}_{i=0}^N$ is known by using the equality

$$\hat{x}_{ML}(i) = f^{i-N}_{(\hat{x}_{ML},N)}(\hat{x}_{ML}(N)). \tag{7.53}$$

The above equality is theoretically equivalent to the equality

$$\hat{x}_{ML}(i) = f^i(\hat{x}_{ML}). \tag{7.54}$$

However, the former equality is more useful in practice because the composed inverse system $f^{-i}_{(x(0),N)}$ is contracting and consequently does not amplify any computer round-off error as is the case with the composed system f^i.

At first glance, it appears that (7.51) is of little value in finding $\hat{x}_{ML}$ and (7.52) is of little value in finding $\hat{x}_{ML}(N)$ since both expressions use the set of affine parameter pairs $\{(\tau(i, \hat{x}_{ML}), \beta(i, \hat{x}_{ML}))\}_{i=0}^N$, knowledge of which apparently requires knowledge of the unknown ML estimate $\hat{x}_{ML}$. In [65, 66, 67], it was shown that for a special class of maps known as generalized tent maps, the affine parameter pairs $\{(\tau(i, \hat{x}_{ML}), \beta(i, \hat{x}_{ML}))\}$ and observations are causally related. For maps in this special class, one can determine each affine parameter pair $(\tau(i, \hat{x}_{ML}), \beta(i, \hat{x}_{ML}))$ from the subset of observations $\{y(j)\}_{j=0}^i$, thereby

allowing recursive, computationally efficient ML orbit point estimation and simultaneous determination of the affine parameter pairs. The method in [65] does not appear to be adaptable to the more general class of MC maps, since the affine parameter pairs and observations are not causally related for these maps in general.

One method to obtain the sequence of affine parameters associated with $\hat{x}_{ML}$ is the computationally intensive, brute-force method of evaluating (7.51) for each possible sequence of affine parameter pairs, determining the most likely orbit segment associated with each sequence, and choosing the sequence of parameter pairs for which the associated most likely orbit segment yields the largest value of the likelihood function. However, the orbit segment obtained with this method which yields the largest value of the likelihood function is in fact the ML orbit segment. What we seek is a computationally simpler estimator for the ML orbit segment. In the next section, we introduce such an estimator; the estimator exploits the relation between MC maps and HMMs discussed earlier in the chapter. The estimator is an optimal ML estimator if the HMM it uses is chosen appropriately. Otherwise, the estimator is a suboptimal ML estimator, but one that is potentially effective nevertheless.

We conclude this section with a subtle, theoretical issue involving ML state estimation and MC maps. An implicit assumption in the discussion thus far has been the existence of $\hat{x}_{ML}$. In fact, if f is not continuous, $\hat{x}_{ML}$ may not exist except in a limiting sense. In particular, consider the situation in which x_{MAX} is not in $A(\hat{x}_{ML})$, and thus the likelihood function attains its maximum among points in $A(\hat{x}_{ML})$ at an endpoint x_e of $A(\hat{x}_{ML})$. However, x_e may not belong to $A(\hat{x}_{ML})$ if f is discontinuous at one of the points of the orbit segment $\{f^i(x_e)\}_{i=0}^N$. In other words, each endpoint of an affine segment domain either belongs to that segment domain or to the domain of the adjacent segment, with the endpoint having two possible images under f depending upon which domain the point belongs to. The implication is that x_e can be the ML estimate of x only if it is in $A(\hat{x}_{ML})$. If $x_e \notin A(\hat{x}_{ML})$, then $\hat{x}_{ML}$ does not exist since the likelihood function is increasing on any infinite sequence of points in $A(\hat{x}_{ML})$ converging to x_e. Whereas the number of discontinuities of f is finite, the probability of it occurring is negligible. Consequently, in the discussion that follows we assume that $\hat{x}_{ML}$ exists and that it is unique, although the results are readily generalized to the case of multiple values of x maximizing the likelihood function.

7.7 Optimal/Suboptimal ML State Estimator for MC Maps

7.7.1 Theoretical Foundation

We now show that for any MC map which gives rise to arbitrarily fine Markov partitions, one can in theory obtain the sequence of affine parameter pairs $\{(\tau(i,\hat{x}_{ML}),\beta(i,\hat{x}_{ML}))\}_{i=0}^{N}$ associated with $\hat{x}_{ML}$ without knowledge of $\hat{x}_{ML}$. As shown in the previous section, if the ML sequence of affine parameter pairs is known, it is straightforward to determine $\hat{x}_{ML}$ and the entire ML orbit segment. The theoretical result presented in this subsection provides the foundation for a practical state estimator introduced in the next subsection. For a given set of observations, the estimator is the optimal ML estimator if a certain HMM used by the estimator is chosen appropriately. The estimator may be a suboptimal ML estimator otherwise.

The theoretical result presented in this subsection and the estimator introduced in the next both exploit the close relation between the likelihoods of orbit segments generated by MC maps for a given set of noisy observations and the likelihoods of state sequences of associated HMMs. To offer some insight into this relation, we consider an MC map f along with a Markov partition $\{I_j\}_{j=0}^{T}$ and its corresponding Markov chain, with S_j denoting the state associated with I_j. Just as we can associate a sequence of affine parameter pairs and affine segment domains with each orbit segment or initial condition, we can associate a sequence of affine parameter pairs $\{(\tau(i,S),\beta(i,S))\}_{i=0}^{N}$ and affine segment domains $\{A(i,S)\}_{i=0}^{N}$ with each state sequence $S \equiv \{S(i)\}_{i=0}^{N}$ of the Markov chain. In particular, if $S(i) = S_j$ and $I_j \subset A_k$, we let $(\tau(i,S),\beta(i,S)) = (\tau_k,\beta_k)$ and $A(i) = A_k$. That is, for each i we associate the affine parameter pair and segment domain for the affine segment of f whose domain contains the partition element associated with $S(i)$.

We let $\{p_{ij}\}_{i,j=1}^{T}$ denote the state transition probabilities of the Markov chain corresponding to the given Markov partition; we let $p(y(i)|S_j)$ denote the output PDF associated with the j^{th} state for either of the HMM models (quantized or unquantized outputs) introduced earlier; and we let $\pi(S_j)$ denote the initial state probability of the j^{th} state. Then, for a given state sequence $S = \{S(i)\}_{i=0}^{N}$ and a given set of observations $Y = \{y(i)\}_{i=0}^{N}$, the joint PDF of the state sequence and observations $p(S,Y)$ is given by

$$p(S,Y) = \pi(S(0))\, p(y(0)|S(0)) \prod_{i=1}^{N} p_{S(i-1),S(i)}\, p(y(i)|S(i)). \tag{7.55}$$

The Viterbi algorithm [74, 75] is a computationally efficient algorithm for finding the state sequence which maximizes $p(S,Y)$, and this state sequence also maximizes $p(S|Y)$, i.e., it is the most probable state sequence for the given observation set. We also have the following expression, which we use later for $P(S) = P(S(0),\cdots,S(N))$, the probability of the state sequence S:

$$P(S) = \pi(S(0)) \prod_{i=1}^{N} p_{S(i-1),S(i)}. \tag{7.56}$$

We now consider the set of state transition pseudo-probabilities $\{q_{ij}\}_{i,j=1}^{T}$ where $q_{ij} = 1$ if $p_{ij} > 0$ and $q_{ij} = 0$ if $p_{ij} = 0$; and we consider the set of initial state pseudo-probabilities $\{\eta(S_j)\}_{i=1}^{N}$, where $\eta(S_j) = 1$ for each j. With these state transition pseudo-probabilities and initial state pseudo-probabilities, we define the joint pseudo-PDF $q(S,Y)$ as

$$q(S,Y) = \eta(S(0))\, p(y(0)|S(0)) \prod_{i=1}^{N} q_{S(i-1),S(i)}\, p(y(i)|S(i)) \tag{7.57}$$

$$= \left\{ \begin{array}{ll} \prod_{i=0}^{N} p(y(i)|S(i)) & \text{if } P(S) > 0 \\ 0 & \text{if } P(S) = 0 \end{array} \right\}. \tag{7.58}$$

As with (7.55), we can use the Viterbi algorithm to efficiently determine the state sequence which maximizes $q(S,Y)$ for a given observation set Y.

We now use these state transition pseudo-probabilities and initial state pseudo-probabilities with the HMM model for quantized orbit points introduced in Section 7.2.2. As in that section, we let H_j denote the value associated with S_j, but we now require that $H_j \in I_j$. In other words, we require that the constant, quantized value associated with each state be a point in the partition element corresponding to that state. For a given state sequence S, we let $H(i)$ denote the quantized value associated with $S(i)$, i.e., $H(i) = H_j$ if $S(i) = S_j$. For this model and the earlier assumption that the observation noise sequence $\{v(i)\}_{i=0}^{N}$ is a Gaussian, white-noise sequence with variance σ^2, the following is true:

$$\log p(y(i)|S(i)) = C(0) - \frac{(y(i) - H(i))^2}{2\,\sigma^2} \tag{7.59}$$

where $C(0)$ is a constant. Therefore, if $P(S) > 0$,

$$\log q(S,Y) = C(N) - \sum_{i=0}^{N} \frac{(y(i) - H(i))^2}{2\sigma^2} \tag{7.60}$$

which is the same expression as $\log p(Y; x(j)$ given by (7.34) if we associate the sequence of state values $\{H(i)\}_{i=0}^{N}$ with the orbit segment $\{f^i(x(0))\}_{i=0}^{N}$. Thus, if the sequence $\{H(i)\}_{i=0}^{N}$ is treated as an orbit segment, then the ML state sequence $S_{ML} = \{S_{ML}(i)\}_{i=0}^{N}$ which we define as the state sequence that maximizes $q(S,Y)$ or equivalently $q(S|Y)$, is also the state sequence with nonzero probability, i.e., $P(S_{ML}) > 0$ which maximizes the log-likelihood function given by (7.34).

In light of the close relation between (7.34) and (7.60), one might expect there to be a relation between the most likely state sequence S_{ML} and the ML orbit segment $\{f^i(\hat{x}_{ML})\}_{i=0}^{N}$. It is plausible that for a given MC map, one can find a sufficiently fine Markov partition such that if $\{I_{ML}(i)\}_{i=0}^{N}$ denotes the sequence of partition elements associated with S_{ML}, then $f^i(\hat{x}_{ML}) \in I_{ML}(i)$ for each i. However, it is unclear how one might prove this relation. In Appendix B, we establish the weaker relation that for a given MC map, a given set of observations, and a sufficiently fine Markov partition, the sequence of affine parameter pairs associated with the most likely state sequence S_{ML} is the same sequence as $\{(\tau(i,\hat{x}_{ML}),\beta(i,\hat{x}_{ML}))\}$, the sequence of affine parameter pairs associated with $\hat{x}_{ML}$. The value of this relation is that for a given MC map and set of observations, if we first select a Markov partition for which this equality in sequences of affine parameter pairs holds and then determine S_{ML} and its associated sequence of affine parameter pairs, we can in theory evaluate $\hat{x}_{ML}$ and the entire ML orbit segment. In light of the discussion in the previous subsection, we can also obtain the ML orbit segment in practice if the map is an EMC map with affine segment slopes having absolute value greater than one and for which the image of each affine segment domain is a union of affine segment domains. In summary and more formally, we have the following result:

Proposition 4: For any MC map f which gives rise to arbitrarily fine Markov partitions and for a given, finite set of observations $Y = \{y(i)\}_{i=0}^{N}$, one can determine the ML sequence of affine parameter pairs $\{(\tau(i,\hat{x}_{ML}),\beta(i,\hat{x}_{ML}))\}$ without knowledge of $\hat{x}_{ML}$, and as a consequence one can in theory determine the ML orbit segment $\{\hat{x}_{ML}(i)\}_{i=0}^{N}$ without testing every sequence of affine parameter pairs.

Proof: (see Appendix B)

Although choosing a Markov partition satisfying the conditions of the proof guarantees that the ML state sequence yields the sequence of affine parameter pairs corresponding to the ML orbit segment, often a partition with fewer elements works as well. A practical iterative method for choosing a sufficiently fine Markov partition is to start with a coarse partition, determine the sequence of affine parameter pairs corresponding to the ML state sequence, and then iteratively select refinements of the partition until the corresponding sequences of affine parameter pairs for the ML state sequences remain the same. In addition, as suggested by the proof of the theorem, as the number of observations N increases, increasingly finer Markov partitions may be needed. For larger values of N, an alternative, possibly suboptimal ML estimator involves using a Markov partition which may not be sufficiently fine to ensure that the sequence of affine parameter pairs associated with the ML state sequence of the corresponding Markov partition is the same sequence as that associated with the ML orbit segment.

7.7.2 The Estimation Algorithm

As noted earlier, the solution of (7.51) yields $\hat{x}_{ML}$ only if it lies in $A(\hat{x}_{ML})$, and it is often impractical to determine $A(\hat{x}_{ML})$ especially for larger values of N. Also noted earlier was the fact that the relation $\hat{x}_{ML}(i) = f^i(\hat{x}_{ML})$ is often of little practical value with EMC maps because of the expansive nature of these maps and computer round-off error in determining $\hat{x}_{ML}$. However, we can circumvent both practical problems with EMC maps for which the affine segment slopes having absolute values greater than one and for which each affine segment domain maps onto a union of affine segment domains. For these maps, we can first find $\hat{x}_{ML}(N)$, generally a far simpler task, and from this calculate the ML orbit segment as $\{f^{-i}_{(\hat{x}_{ML},N)}(\hat{x}_{ML}(N)\}_{i=0}^{N}$. The two EMC maps used for the examples in the next subsection both satisfy these constraints.

Fusing these practical considerations with the theoretical result presented in the previous subsection leads to the following following practical algorithm for estimating the ML orbit segment based on a set of observations Y for any EMC map f which gives rise to arbitrarily fine Markov partitions and for which the affine segment slopes have absolute values greater than one and for which each affine segment domain maps onto a union of affine segment domains.

Optimal/Suboptimal ML State Estimator

1. Given an $(N+1)$-point observation set $Y(0,N)$ and a Markov partition for f, find the ML state sequence for the quantized-output HMM model associated with this partition in which the quantized output value for each state is given by any point in the corresponding partition element and which uses the state transition pseudo-probabilities and initial state pseudo-probabilities defined in the previous subsection. Let S_{ML} denote this state sequence.

2. Given the ML state sequence, find the associated sequences of affine parameter pairs $\{\tau(i,S_{ML}),\beta(i,S_{ML}\}_{i=0}^{N}$ and affine segment domains $\{A(i,S_{ML})\}_{i=0}^{N}$. If the Markov partition is sufficiently fine so as to satisfy the sufficient condition specified in the proof of Proposition 4, the sequences will be the same as those associated with the ML orbit segment. If not, the sequences may not be the same as those associated with the ML orbit segment, and the algorithm may not be optimal.

3. Evaluate (7.52) for $x_{MAX}(N)$ using the sequence of affine parameter pairs associated with the ML state sequence.

4. If $x_{MAX}(N) \in A(N,S_{ML})$, set $\hat{x}_{ML}(N)$ equal to $x_{MAX}(N)$. If not, set $\hat{x}_{ML}(N)$ equal to the endpoint of $A(N,S_{ML})$ for which $\log p(Y;x(N))$ has the larger value.

5. Let $\{f^{-i}_{(S_{ML},N)}\}_{i=0}^{N}$ denote the function implicitly defined by $\{\tau(i,S_{ML}),\beta(i,S_{ML}\}_{i=0}^{N}$ (analogous to the function $\{f^{-i}_{(\hat{x}_{ML},N)}\}_{i=0}^{N}$ defined earlier). In particular, we have the following for $i=1$ and $i=2$:

$$x(N-1) = f^{-1}_{(S_{ML},N)}(x(N)) \equiv \frac{x(N)-\beta(N-1,S_{ML})}{\tau(N-1,S_{ML})} \tag{7.61}$$

$$x(N-2) = f^{-2}_{(S_{ML},N)}(x(N)) \equiv \frac{x(N-1)-\beta(N-2,S_{ML})}{\tau(N-2,S_{ML})}. \tag{7.62}$$

6. Calculate the reverse orbit segment $\{\hat{x}_{ML}(N-i) = f^{-i}_{(S_{ML},N)}(\hat{x}_{ML}(N))\}$. This orbit segment is the ML orbit segment if the Markov partition is sufficiently fine.

7.7.3 Estimation Examples

Figures 7-17 (a) and (b) depict the performance results obtained by applying the ML estimator to the maps shown in Figures 7-16 (a) and (b), respectively. Both maps were used for

examples earlier in the chapter. Each plotted SNR gain reflects the average improvement

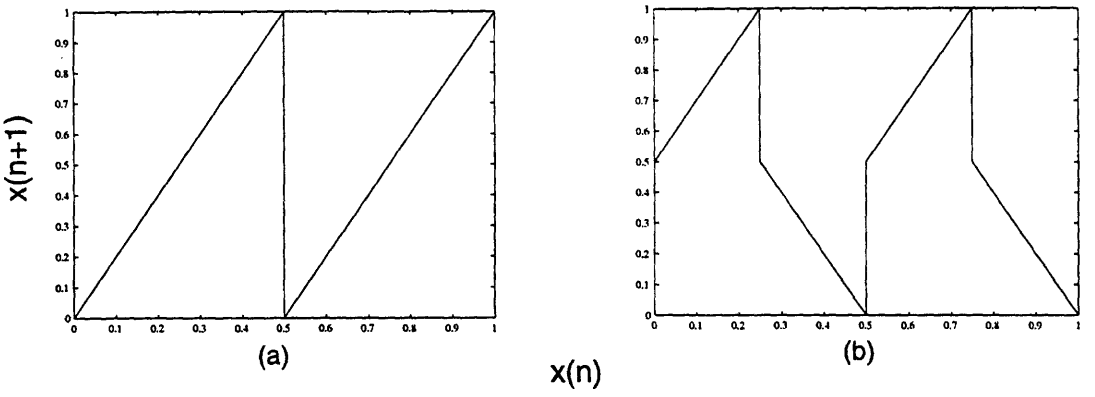

Figure 7-16: Two EMC maps. (a) g_1: (b) g_2.

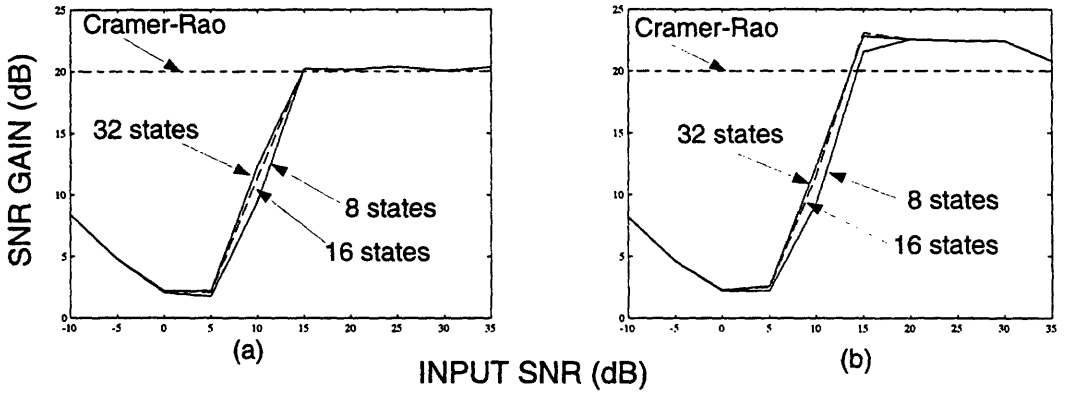

Figure 7-17: Performance results for estimating 100-point orbit segment. (a) g_1: (b) g_2.

in SNR obtained by estimating each point of a 100-point orbit segment. More precisely, for each input SNR, 1000 independent observation sets were generated with the same 100-point orbit segment. The estimator was applied to each of the observation sets, and the average SNR improvement in estimating the 100 orbit segment points was calculated for each observation set. The 1000 values of average SNR improvement were then averaged and used in the figure as the plotted SNR gain for the corresponding input SNR.

The curves are parameterized by the number of states in the HMM model used to estimate the sequence of affine parameters associated with the ML orbit segment. Also shown in the figures is the upper bound on the SNR gain provided by the Cramer-Rao bound for *unbiased* estimators. Using the more general results in [76], one can show that the Cramer-Rao bound for fixed-interval smoothing with one-dimensional maps is given by σ^2/N where N is the number of orbit segment points and σ^2 is the observation noise

variance.

The figures suggest that the ML estimator is superoptimal in the sense that the SNR gain exceeds the Cramer-Rao bound at some input SNRs. However, such is not the case as the ML estimator is biased for these maps, and thus the Cramer-Rao bound for unbiased estimators is only applicable to the estimator in an asymptotic sense as the input SNR tends to infinity. Although deriving an analytic expression for the bias appears to be an intractable problem, we can examine its behavior in simulations. Figures 7-18 (a) and (b) depict the absolute value of the bias at each orbit point for the results shown in Figures 7-17 (a) and (b), respectively. The larger bias values at the end of the orbit segment is

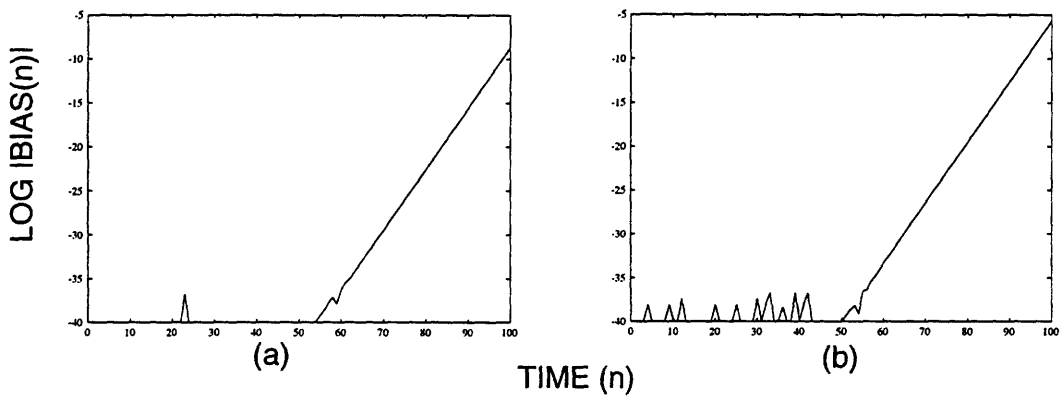

Figure 7-18: Estimator bias for results in Figure 7-17. (a) g_1: (b) g_2.

understandable in light of the fact that the squared estimation error is largest at the end of the segment. The performance results for both maps exhibit a threshold effect with significant SNR gain for input SNRs above a threshold and mediocre or negligible SNR gain for input SNRs below the threshold. Such an effect is typical of ML estimators for nonlinear estimation problems, with the threshold indicating the input SNR value below which nonlinearities strongly influence the likelihood function. Large SNR gains at larger input SNRs is not peculiar to the chosen orbit segment as suggested by Figures 7-19 (a) and (b), which depict the average of the performance results for 100 randomly chosen 100-point orbit segments.

Because both maps are noninvertible, one-dimensional, and exhibit sensitive dependence on initial conditions, estimation accuracy improves as the number of observations at future times increases (as indicated by the performance bound analysis in Chapter 5). As a result and as noted above, at higher input SNRs the estimation errors associated with the final points of an orbit segment dominate the average of squared estimation errors for all orbit

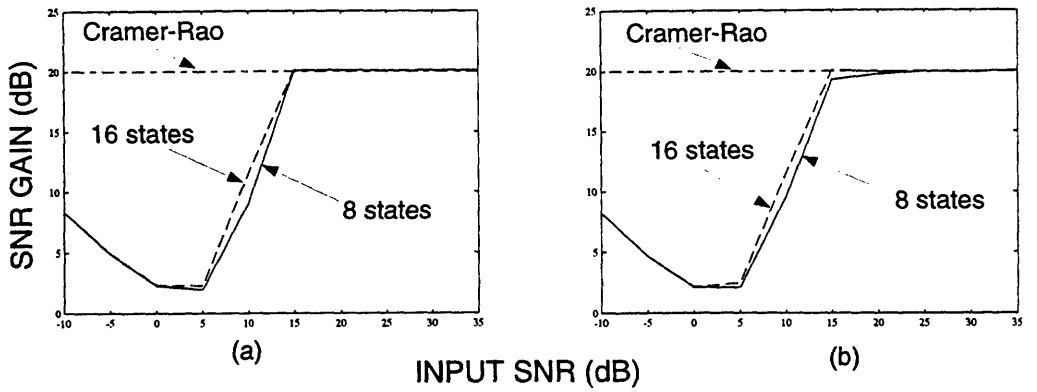

Figure 7-19: Average of performance results for estimating 100 randomly chosen 100-point orbit segments. (a) g_1: (b) g_2.

points. Figures 7-20 (a) and (b) depict the performance results for the same conditions as used for the results in Figures 7-17 (a) and (b), but with the estimation errors for the last 30 points of the estimated 100-point orbit segment, omitted from the SNR gain calculation. A

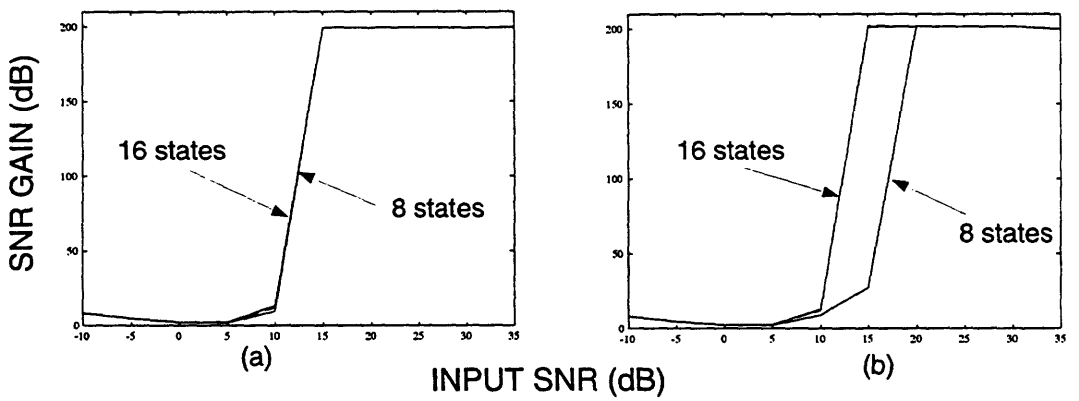

Figure 7-20: Performance results for estimating 100-point orbit segments with estimates of last 30 points not used in gain calculation. (a) g_1: (b) g_2.

comparison of corresponding pairs of figures reveals the dominant role that the estimation error for these omitted points has on the average squared estimation error.

Finally, the question arises as to the value of using an HMM model to estimate the sequence of affine parameter pairs associated with the ML orbit segment. Figures 7-21 (a) and (b) depict the performance results obtained by using the sequence of affine parameter pairs associated with the observation set Y. That is, for each i the parameter pair (τ_j, β_j) for which $y(i) \in A_j$ was used for $(\tau(i, \hat{x}_{ML}), \beta(i, \hat{x}_{ML}))$. The poor performance results suggest the necessity and value of using an HMM to estimate the sequence of affine parameter pairs associated with the ML orbit segment.

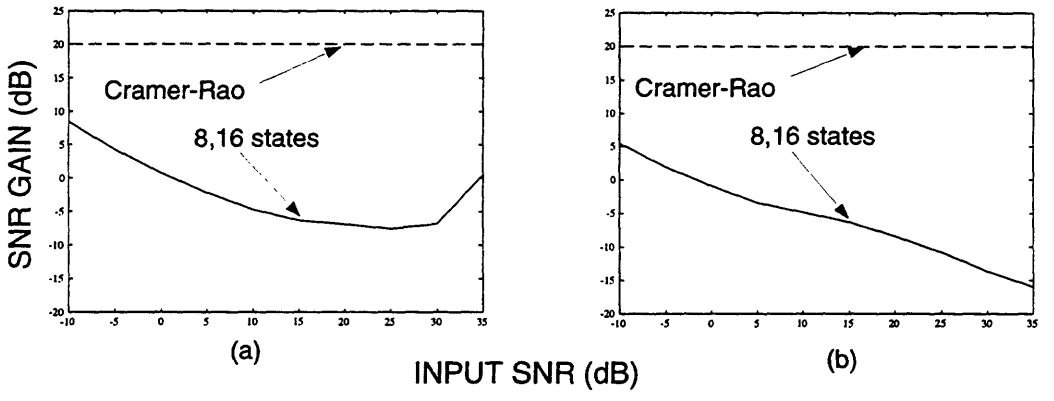

Figure 7-21: Performance results for estimating 100-point orbit segment with sequence of affine parameter pairs associated with the observation set used as the estimated sequence associated with the ML orbit segment. (a) f_1: (b) f_2.

Chapter 8

Conclusions

8.1 Summary and Contributions

This thesis has dealt with the analysis and synthesis of chaotic maps and time-sampled chaotic flows, with a focus on the problems and issues that arise with noise-corrupted orbit segments generated by these systems. Both dissipative systems and nondissipative systems have been considered, with both types of systems considered in the context of analysis and the latter type also considered in the context of synthesis. With respect to dissipative systems, three suboptimal, probabilistic state estimation algorithms—an ML estimator based on grid search, a local MMSE estimator based on extended Kalman smoothing, and a global MMSE estimator based on a finite-sum approximation to the conditional mean integral—have been introduced and their performance experimentally assessed on three different problem scenarios: known system dynamics, unknown system dynamics but availability of a noise-free reference orbit, unknown system dynamics and no availability of a noise-free reference orbit. Both the ML and local MMSE estimators exploit the topological transitivity of dissipative, chaotic systems when restricted to their steady-state attractors, and both estimators are potentially effective for all three problem scenarios. The global MMSE estimator exploits the existence of an ergodic measure on a steady-state chaotic attractor, which allows the substitution of an infinite summation for the integral defining the conditional mean which yields the optimal MMSE state estimator. One feature of the global MMSE estimator is that it converges to the optimal MMSE state estimator as the number of terms in the finite summation defining the estimator goes to infinity.

The assumed determinism in the system dynamics facilitates the derivation of upper

bounds—Cramer-Rao, Barankin, and Weiss-Weinstein—on state estimator performance. These bounds have been derived for the state estimation problem of interest in the thesis, and their behavior has been experimentally analyzed on two, dissipative, chaotic systems: the Henon map and time-sampled Lorenz flow. The Cramer-Rao and Barankin bounds have been shown to provide potentially useful information on the influence of fundamental properties of dissipative, chaotic diffeomorphisms—positive Lyapunov exponents and boundedness of attractors—on achievable state estimator performance, when the unknown state is nonrandom. In contrast, the random Cramer-Rao and Weiss-Weinstein bounds have been shown to be of limited value in the context of initial condition estimation with dissipative, chaotic diffeomorphisms, when the unknown initial condition is a random vector.

With respect to nondissipative maps, the thesis has considered a class of piecewise linear unit-interval maps, members of which give rise to finite-state, homogeneous Markov chains. The thesis has presented known properties of these maps, established additional properties, and explored the potential value of these maps as generators of signals for practical applications. A close relation between noise-corrupted orbit segments generated by the maps and hidden Markov models has been established, and this relation has been exploited in practical, optimal and suboptimal algorithms for detection, parameter estimation, and state estimation with the maps.

This thesis has made two, principal contributions to the research community. First, it has established a rigorous, probabilistic foundation for the problem of state estimation with deterministic, dissipative, chaotic diffeomorphisms. In particular, the thesis has shown how the existence of invariant measures on steady-state chaotic attractors facilitates practical state estimation with dissipative, chaotic systems. In addition, the thesis has shown the value of the Cramer-Rao and Barankin bounds for assessing the influence of fundamental properties of chaotic systems including the existence of positive Lyapunov exponents, boundedness of attractors, and system invertibility or noninvertibility on theoretically achievable state estimator performance with these systems. Finally, the thesis has exposed the limitations of performance bounds for estimators of random parameters when applied to dissipative, chaotic systems.

The second, principal contribution of the thesis involves its consideration of MC maps. In particular, by assimilating known and establishing additional properties of these maps, the thesis has identified a potentially useful and versatile source of signal generators. In

addition, the detection, parameter estimation, and state estimation algorithms introduced in the thesis for use with these maps may prove to be of value in applications involving the maps. These algorithms may also provide useful insight into the design of optimal, effective, and robust algorithms for detection, parameter estimation, and state estimation with other classes of chaotic systems such as dissipative maps and flows.

8.2 Suggested Future Research

This thesis has either partially or totally resolved several research problems involving chaotic systems; and as a result, several new research problems have emerged. Consequently, the analysis, algorithms, and experimental results presented in this thesis implicitly suggest a number of interesting, challenging, and potentially fruitful topics for future research.

With respect to state estimation with dissipative maps, the state estimation algorithms and performance bound derivations introduced in Chapters 4 and 5 may have shed new light on the problem of state estimation with chaos, but optimal, practical, probabilistic state estimation with chaos remains an elusive goal. It would be useful to explore methods to refine the approximate ML and MMSE state estimators introduced in Chapter 4 so that the simplifying, practical, heuristic elements of the estimators are avoided. Alternatively, it would be useful to establish a theoretical justification for these heuristic elements and formulate theoretically motivated guidelines for choosing values for the various parameters which these elements introduce into the estimators. In the pursuit of these tasks, it might be useful to replace the deterministic state equation used throughout the thesis by a sequence of noise-driven state equations with the deterministic equation as the limit.

An additional topic for future research arises from the similarities of the global MMSE state estimator introduced in Chapter 4 with the hidden Markov modeling (HMM) estimator introduced in [63], which also is a global estimator. Each of the estimators has strengths and weaknesses over the other, and computer experiments suggests that each is a potentially effective state estimator with dissipative, chaotic maps. A useful research task would involve merging the two estimators in an effort to create a hybrid state estimator that retains the strengths and avoids the weaknesses of the individual estimators. Useful insight into both estimators might be obtained by studying them in conjunction with the class of MC maps introduced in Chapter 6. In fact, as noted at the beginning of that chapter,

our original purpose in studying MC maps was to better understand the theoretical and practical properties of the HMM estimator. A related task, not considered in the thesis, would involve extending the global MMSE estimator to the self-cleaning problem scenario, that is, the state estimation scenario in which the system dynamics are unknown and a noise-free reference orbit segment is not available. Computer experiments suggest that the estimator in its present form is not useful for this scenario.

With respect to state estimator performance bounds, several questions remain unresolved and several new research problems have emerged thereby suggesting a number of challenging, potentially beneficial research tasks. Perhaps the most important and useful of these tasks would be the development of new state estimator performance bounds for use with dissipative, chaotic diffeomorphisms when the unknown state is a random vector. The importance of this task stems from the experimental results presented in Chapter 5 which graphically indicated that existing performance bounds for estimators of random parameters have limited value with dissipative, chaotic systems. In addition, many of the fundamental conclusions of the chapter involving the influence of Lyapunov exponents, attractor boundedness, and system invertibility on achievable state estimation performance, are strictly applicable only to unbiased state estimators. It appears that this restriction on the bias is an intrinsic aspect of all existing performance bounds for nonrandom-parameter estimators which is avoidable only by using estimator-dependent bounds or by treating the unknown parameter as random. In light of this, it appears that extending the fundamental conclusions of the chapter to arbitrary state estimators, both unbiased and biased, requires consideration of performance bounds for estimators of random parameters, thereby underscoring the need for new, effective performance bounds for the problem of state estimation with chaotic systems when the unknown state is random.

Several unresolved or unexplored issues concerning the unit-interval maps considered in Chapters 6 and 7 remain; a number of worthwhile, potentially fruitful research tasks should be undertaken to resolve these issues. For example, as suggested in Chapter 6, MC maps might be useful signal generators in light of the rich set of properties these maps exhibit. It would be useful to ascertain the strengths and weaknesses of using MC maps and maps derived from them over other techniques for generating Markov chains and processes with specified stationary PDFs. It would also be useful to identify specific, practical applications in which MC maps might be used, such as secure communication, and to assess the value

of using MC maps for these applications. An unrelated research task involves the detection algorithm introduced in Chapter 7 for discriminating among EMC maps based on noise-corrupted, unquantized orbit segments. Although the algorithm is suboptimal, it appears that the algorithm converges in an appropriately defined sense to the optimal, minimum probability-of-error detector, as the Markov partition used by the detector becomes increasingly fine. It would be useful to formally and conclusively establish convergence properties of the detector.

Only a relatively small class of deterministic, unit interval maps was considered in Chapters 6 and 7. However, this class of maps is dense (in an appropriately defined sense) in a much larger set of unit-interval maps [14, 46]. It would be useful to determine if the detection and estimation algorithms introduced in Chapter 7 can be extended to this larger set of maps. Similarly, it would be useful to determine if the properties of MC maps are shared by other classes of deterministic maps and flows, and if so, if the detection and estimation algorithms introduced in the thesis can be extended to these other classes of deterministic systems as well.

Appendix A

Performance Bound Equations

A.1 Cramer-Rao Bound

In this section, we derive the Cramer-Rao bound for the problem scenario of interest in Chapter 5 when the unknown state to be estimated is nonrandom. The derivation of the Cramer-Rao bound for deterministic, nonlinear systems has appeared elsewhere (see e.g., [85]); we include it here for completeness and for a more general problem scenario, one involving a variable number of observations occurring at past and future times, than has been considered in the past. In particular, we derive the Fisher information matrix $\boldsymbol{J}(\boldsymbol{x}_{n_0})$, the inverse of which provides the Cramer-Rao bound on the error covariance matrix of any unbiased estimator for $\boldsymbol{x}(n_0)$ when $\boldsymbol{x}(n_0) = \boldsymbol{x}_{n_0}$, where $\boldsymbol{x}_{n_0}$ is the actual value of the state at time n_0.

For the DTS/DTO model given by (5.1) and (5.2) and with the slightly more general assumption that the observation noise sequence is white, but not necessarily Gaussian with the PDF of each sequence element given by $p_V(v)$, the PDF $p(\boldsymbol{y}(i); \boldsymbol{x}(i))$ is given by

$$p(\boldsymbol{y}(i); \boldsymbol{x}(i)) = p_V(\boldsymbol{y}(i) - \boldsymbol{h}(\boldsymbol{x}(i))) \tag{A.1}$$

and since $\boldsymbol{x}(i) = \boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0))$ and the noise is white, it follows that the likelihood function $p(Y; \boldsymbol{x}(n_0))$ is given by

$$p(Y; \boldsymbol{x}(n_0)) = \prod_{i=M}^{N} p_V(\boldsymbol{y}(i) - \boldsymbol{h}(\boldsymbol{x}(i))) \tag{A.2}$$

$$= \prod_{i=M}^{N} p_V(\boldsymbol{y}(i) - \boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))), \tag{A.3}$$

and therefore

$$\log p(Y; \boldsymbol{x}(n_0)) = \sum_{i=M}^{N} \log p_V(\boldsymbol{y}(i) - \boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))). \tag{A.4}$$

To establish the Cramer-Rao bound for this problem scenario, we need to determine the Fisher information matrix $\boldsymbol{J}(\boldsymbol{x}_{n_0})$, the general form of which is given by

$$\boldsymbol{J}(\boldsymbol{x}_{n_0}) = E_{Y;\boldsymbol{x}_0}\left\{\boldsymbol{D}^T_{\boldsymbol{x}(n_0)}\left\{\log p(Y; \boldsymbol{x}_{n_0})\right\} \boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\log p(Y; \boldsymbol{x}_{n_0})\right\}\right\} \tag{A.5}$$

where $\boldsymbol{D}^T_{\boldsymbol{x}(n_0)}\{\log p(Y; \boldsymbol{x}_{n_0})\}$ denotes the derivative of $\log p(Y; \boldsymbol{x}(n_0))$ taken with respect to $\boldsymbol{x}(n_0)$ and evaluated at $\boldsymbol{x}_{n_0}$. Applying the chain rule of vector differentiation to (A.4) yields

$$\boldsymbol{D}_{\boldsymbol{x}(n_0)}\{\log p(Y; \boldsymbol{x}_{n_0})\} = \sum_{i=M}^{N} \boldsymbol{D}\{\log p_V(\boldsymbol{v}(i))\} \boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}_{n_0}))\right\} \tag{A.6}$$

where $\boldsymbol{v}(i) = \boldsymbol{y}(i) - \boldsymbol{h}(\boldsymbol{f}^{i-n}(\boldsymbol{x}(n_0)))$. Note that $\boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right\}$ is expressible as a product of the derivative of $\boldsymbol{f}$ or $\boldsymbol{f}^{-1}$, specifically

$$\boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right\} = \boldsymbol{D}\left\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right\} \boldsymbol{D}\{\boldsymbol{f}(\boldsymbol{x}(i-1))\} \cdots \boldsymbol{D}\{\boldsymbol{f}(\boldsymbol{x}(n_0))\} \tag{A.7}$$

for $i \geq n_0 + 2$, as

$$\boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\boldsymbol{h}(\boldsymbol{f}^{i-n}(\boldsymbol{x}(n_0)))\right\} = \boldsymbol{D}\left\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}(n_0)))\right\} \boldsymbol{D}\{\boldsymbol{f}^{-1}(\boldsymbol{x}(i+1))\} \cdots \boldsymbol{D}\{\boldsymbol{f}^{-1}(\boldsymbol{x}(n_0))\} \tag{A.8}$$

for $i \leq n_0 - 2$, and with analogous expressions for $n_0 - 1 \leq i \leq n_0 + 1$.

With the appropriate substitutions, the Fisher information matrix becomes

$$\boldsymbol{J}(\boldsymbol{x}_{n_0}) = \sum_{i=M}^{N} \sum_{j=M}^{N} \boldsymbol{D}^T_{\boldsymbol{x}(n_0)}\left\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}_{n_0}))\right\} \boldsymbol{Q}(i,j) \boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\boldsymbol{h}(\boldsymbol{f}^{j-n_0}(\boldsymbol{x}_{n_0}))\right\} \tag{A.9}$$

where the matrix $\boldsymbol{Q}(i,j)$ is given by

$$\boldsymbol{Q}(i,j) = E_{Y;\boldsymbol{x}_0}\left\{\boldsymbol{D}^T\left\{\log p_V(\boldsymbol{v}(i))\right\}\,\boldsymbol{D}\left\{\log p_V(\boldsymbol{v}(j))\right\}\right\}. \tag{A.10}$$

This decomposition effectively decouples the dependence of $\boldsymbol{J}(\boldsymbol{x}_{n_0})$ on the statistics of the noise $\boldsymbol{v}(n)$ and the dynamics of the system $\boldsymbol{f}$, with the noise statistics reflected in the matrices $\boldsymbol{Q}(i,j)$. When the observation noise is also Gaussian with covariance matrix $\boldsymbol{R}$,

$$\boldsymbol{Q}(i,j) = \boldsymbol{R}\,\delta_{i,j} \tag{A.11}$$

where $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise, and $\boldsymbol{J}(\boldsymbol{x}_{n_0})$ reduces to the following:

$$\boldsymbol{J}(\boldsymbol{x}_{n_0}) = \sum_{i=M}^{N} D^T_{\boldsymbol{x}(n_0)}\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}_{n_0}))\}\boldsymbol{R}^{-1}D_{\boldsymbol{x}(n_0)}\{\boldsymbol{h}(\boldsymbol{f}^{i-n_0}(\boldsymbol{x}_{n_0}))\}. \tag{A.12}$$

A.2 Barankin Bound for Vector-Valued Parameters

The Barankin bound for unbiased estimators of scalar-valued parameters given by (5.19) has a counterpart, derived in [58, 59], for unbiased estimators of vector-valued parameters. For the problem scenario of interest in Chapter 5 in which we seek to bound $\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0))$, the error covariance matrix for the unbiased estimator $\hat{\boldsymbol{x}}(n_0)$ for $\boldsymbol{x}_{n_0}$ given the observation set Y, the general form of the Barankin bound is the following:

$$\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0)) \geq \boldsymbol{X}\,\boldsymbol{Q}^{-1}\,\boldsymbol{X}^T. \tag{A.13}$$

In this equation, $\boldsymbol{X}$ is the $\mathcal{N} \times m$-matrix given by

$$\boldsymbol{X} = [\boldsymbol{x}_1(n_0) - \boldsymbol{x}_{n_0}, \boldsymbol{x}_2(n_0) - \boldsymbol{x}_{n_0}, \cdots, \boldsymbol{x}_m(n_0) - \boldsymbol{x}_{n_0}], \tag{A.14}$$

where $\{\boldsymbol{x}_i(n_0)\}_{i=1}^m$, the test points, denote values of $\boldsymbol{x}(n_0)$ other than $\boldsymbol{x}_{n_0}$. Also, $\boldsymbol{Q}$ is an $m \times m$-matrix with ij^{th} element q_{ij} given by

$$q_{ij} \;=\; E_{Y;\boldsymbol{x}_{n_0}}\left\{L(Y;\boldsymbol{x}_i(n_0),\boldsymbol{x}_{n_0})\,L(Y;\boldsymbol{x}_j(n_0),\boldsymbol{x}_{n_0})\right\} \tag{A.15}$$

$$= \int L(Y;\boldsymbol{x}_i(n_0),\boldsymbol{x}_{n_0})\,L(Y;\boldsymbol{x}_j(n_0),\boldsymbol{x}_{n_0})\,p(Y;\boldsymbol{x}_{n_0})\,dY \tag{A.16}$$

where $L(Y;\boldsymbol{x}_k(n_0),\boldsymbol{x}_{n_0})$ is the likelihood-ratio given by

$$L(Y;\boldsymbol{x}_k(n_0),\boldsymbol{x}_{n_0}) = \frac{p(Y;\boldsymbol{x}_k(n_0))}{p(Y;\boldsymbol{x}_{n_0})}. \tag{A.17}$$

As shown in [59], by appropriately augmenting the original set of test points, one can derive a restricted form of the above bound which is expressible as the sum of two components—one the inverse of the Fisher information matrix, and the other a positive semidefinite matrix which depends on the test points. In particular, this restricted form of the Barankin bound can be expressed as

$$\boldsymbol{P}(\hat{\boldsymbol{x}}(n_0)) \geq \boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0}) + (\boldsymbol{X} - \boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})\,\boldsymbol{A})\boldsymbol{\Delta}^{-1}(\boldsymbol{X} - \boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})\,\boldsymbol{A})^T \tag{A.18}$$

where

$$\boldsymbol{\Delta} = \boldsymbol{B} - \boldsymbol{A}^T\boldsymbol{J}^{-1}(\boldsymbol{x}_{n_0})\,\boldsymbol{A} \tag{A.19}$$

$$\boldsymbol{J}(\boldsymbol{x}_{n_0}) = \text{the Fisher Information Matrix}$$

$$= E_{Y;\boldsymbol{x}_0}\left\{\boldsymbol{D}^T_{\boldsymbol{x}(n_0)}\left\{\log p(Y;\boldsymbol{x}_{n_0})\right\}\boldsymbol{D}_{\boldsymbol{x}(n_0)}\left\{\log p(Y;\boldsymbol{x}_{n_0})\right\}\right\} \tag{A.20}$$

$$A_{ij} = E_{Y;\boldsymbol{x}_0}\left\{\frac{\partial \log p(Y;\boldsymbol{x}_{n_0})}{\partial x(n_0,i)}L(Y;\boldsymbol{x}_j(n_0),\boldsymbol{x}_{n_0})\right\} \tag{A.21}$$

$$i = 1,2,\cdots,\mathcal{N};\ j = 1,2,\cdots,m$$

$$B_{ij} = E_{Y;\boldsymbol{x}_0}\left\{L(Y;\boldsymbol{x}_i(n_0),\boldsymbol{x}_{n_0})\,L(Y;\boldsymbol{x}_j(n_0),\boldsymbol{x}_{n_0})\right\} \tag{A.22}$$

$$i,j = 1,2,\cdots,m$$

where $\boldsymbol{x}(n_0) = [x(n_0,1), x(n_0,2),\cdots,x(n_0,\mathcal{N})]$. Since $\boldsymbol{J}(\boldsymbol{x}_{n_0})$ is the Fisher information matrix and the second term in (A.18) is positive semidefinite (as noted in [59]), this special form of the Barankin bound always provides a tighter bound on the error covariance matrix than the Cramer-Rao bound.

For the problem scenario of interest in Chapter 5 for the special case in which the output transformation $\boldsymbol{h}$ is the identity operator, a straightforward derivation yields the following

values for the elements of $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\boldsymbol{A}_{\cdot,j} = \left\{ \sum_{k=M}^{N} [\boldsymbol{f}^{k-n_0}(\boldsymbol{x}_j(n_0)) - \boldsymbol{f}^{k-n_0}(\boldsymbol{x}_{n_0})]^T \boldsymbol{R}^{-1} D_{\boldsymbol{x}(n_0)} \{\boldsymbol{f}^{k-n_0}(\boldsymbol{x}_{n_0})\} \right\}^T \quad (A.23)$$

$$j = 1, 2, \cdots, m$$

$$B_{ij} = \exp \left\{ \sum_{k=M}^{N} [\boldsymbol{f}^{k-n_0}(\boldsymbol{x}_i(n_0)) - \boldsymbol{f}^{k-n_0}(\boldsymbol{x}_{n_0})]^T \times \boldsymbol{R}^{-1} [\boldsymbol{f}^{k-n_0}(\boldsymbol{x}_j(n_0)) - \boldsymbol{f}^{k-n_0}(\boldsymbol{x}_{n_0})] \right\} \quad (A.24)$$

$$i, j = 1, 2 \cdots, m$$

where $\boldsymbol{A}_{\cdot,j}$ denotes the j^{th} column of $\boldsymbol{A}$.

A.3 Weiss-Weinstein Bound for Vector-Valued Parameters

Analogous to the Barankin bound, the Weiss-Weinstein bound for estimators of scalar-valued, random parameters given by (5.4.2) has a counterpart derived in [90] for estimators of vector-valued, random parameters. For the problem scenario of interest in Chapter 5 in which we seek to bound $\boldsymbol{P}_R(\hat{\boldsymbol{x}}(0))$, the error covariance matrix for estimators of the random initial condition $\boldsymbol{x}(0)$ given the observation set Y, the general form of the Weiss-Weinstein bound is the following:

$$\boldsymbol{P}_R(\hat{\boldsymbol{x}}(0)) \geq \boldsymbol{Z}\, \boldsymbol{W}^{-1}\, \boldsymbol{Z}^T. \quad (A.25)$$

In this equation, $\boldsymbol{Z}$ is the $\mathcal{N} \times m$-matrix given by

$$\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_m], \quad (A.26)$$

with each $\boldsymbol{z}_i$ an $\mathcal{N}$-dimensional vector and $\boldsymbol{W}$ is an $m \times m$-matrix with ij^{th} element w_{ij} given by

$$w_{ij} = E_{Y,\boldsymbol{x}(0)} \{\mathcal{L}_i(Y, \boldsymbol{x}(0))\, \mathcal{L}_j(Y, \boldsymbol{x}(0))\} \quad (A.27)$$

where

$$\mathcal{L}_k(Y, \boldsymbol{x}(0)) = \frac{L^{s_k}(Y, \boldsymbol{x}(0) + \boldsymbol{z}_k, \boldsymbol{x}(0)) - L^{1-s_k}(Y, \boldsymbol{x}(0) - \boldsymbol{z}_k, \boldsymbol{x}(0))}{L^{1-s_k}(Y, \boldsymbol{x}(0) - \boldsymbol{z}_k, \boldsymbol{x}(0))} \quad (A.28)$$

$$L(Y, \boldsymbol{x}_1(0), \boldsymbol{x}_2(0)) = \frac{p(Y, \boldsymbol{x}_1(0))}{p(Y, \boldsymbol{x}_2(0))} \quad (A.29)$$

and where $0 < s_k < 1$. In the equations that follow, we set each s_k equal to .5 as suggested in [90].

As with the Barankin bound although never shown in the literature, with the appropriate selection of test offsets, one can derive a restricted form of the Weiss-Weinstein bound that is expressible as the sum of two components—one the inverse of the Fisher information matrix for random parameters, and the other a function of test offsets, the number of observations, and the observation noise intensity. To do so, one augments the given set of test offsets with $\mathcal{N}$ additional offsets $\{\boldsymbol{z}_i\}_{i=m+1}^{m+\mathcal{N}}$, with $\boldsymbol{z}_{m+j} = z\,\boldsymbol{e}_j$ for $j = 1, 2 \cdots, \mathcal{N}$, where z is a scalar and $\boldsymbol{e}_j$ is the j^{th} unit vector in $\mathcal{R}^{\mathcal{N}}$. In the limit $z \to 0$, the Weiss-Weinstein bound approaches the following:

$$\boldsymbol{P}_R(\hat{\boldsymbol{x}}(0)) \geq \boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0)) + (\boldsymbol{Z} - \boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0))\,\boldsymbol{A})\boldsymbol{\Delta}^{-1}(\boldsymbol{Z} - \boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0))\,\boldsymbol{A})^T \tag{A.30}$$

where

$$\begin{aligned}
\boldsymbol{J}_R(\boldsymbol{x}(0)) &= \text{the Fisher Information Matrix for } \boldsymbol{x}(0) \\
&= E_{Y,\boldsymbol{x}(0)}\left\{\boldsymbol{D}^T_{\boldsymbol{x}(0)}\left\{\log p(Y, \boldsymbol{x}(0))\right\}\boldsymbol{D}_{\boldsymbol{x}(0)}\left\{\log p(Y, \boldsymbol{x}(0))\right\}\right\} && \text{(A.31)} \\
\boldsymbol{Z} &= \text{the } \mathcal{N} \times m\text{-matrix of test offsets} \\
&= [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_m] && \text{(A.32)} \\
\boldsymbol{\Delta} &= \boldsymbol{B} - \boldsymbol{A}^T\boldsymbol{J}_R{}^{-1}(\boldsymbol{x}(0))\,\boldsymbol{A} && \text{(A.33)} \\
A_{ij} &= E_{Y,\boldsymbol{x}(0)}\left\{\frac{\partial \log p(Y, \boldsymbol{x}(0))}{\partial x_i(0)}\,\mathcal{L}_j(Y, \boldsymbol{x}(0))\right\} && \text{(A.34)} \\
&\quad i = 1, 2, \cdots, \mathcal{N};\ j = 1, 2, \cdots, m \\
B_{ij} &= E_{Y,\boldsymbol{x}(0)}\left\{\mathcal{L}_i(Y, \boldsymbol{x}(0))\,\mathcal{L}_j(Y, \boldsymbol{x}(0))\right\} && \text{(A.35)} \\
&\quad i, j = 1, 2, \cdots, m
\end{aligned}$$

where $\boldsymbol{x}(0) = [x_1(0), x_2(0), \cdots, x_{\mathcal{N}}(0)]$.

For the problem scenario of interest in Chapter 5 for the special case in which the observation noise covariance matrix $\boldsymbol{R}$ equals $\sigma^2\,\boldsymbol{I}_{\mathcal{N}}$ and the output transformation $\boldsymbol{h}$ is the identity operator, a straightforward derivation yields the following values for the elements

of $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\boldsymbol{A}_{\cdot,j} = E_{\boldsymbol{x}(0)}\left\{\frac{\mathcal{N}_{\mathcal{A}}(j)}{\mathcal{D}_{\mathcal{A}}(j)}\right\}, \quad j = 1, 2, \cdots, m \tag{A.36}$$

$$\begin{aligned}
\mathcal{N}_{\mathcal{A}}(j) &= \left[\frac{p(\boldsymbol{x}(0)+\boldsymbol{z}_j)}{p(\boldsymbol{x}(0))}\right]^{.5} \exp\left[-\frac{1}{8\sigma^2}\left\{\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)+\boldsymbol{z}_j)-\boldsymbol{f}^k(\boldsymbol{x}(0))]^2\right\}\right] \\
&\times \left\{\left[\sum_{k=0}^{N}\frac{\left[\boldsymbol{f}^k(\boldsymbol{x}(0)+\boldsymbol{z}_j)-\boldsymbol{f}^k(\boldsymbol{x}(0))\right]^T}{2\sigma^2}\boldsymbol{D}\{\boldsymbol{f}^k(\boldsymbol{x}(0))\}\right] + \frac{[\boldsymbol{x}_0-\boldsymbol{x}(0)]^T}{\gamma^2}\right\}^T \\
&- \left[\frac{p(\boldsymbol{x}(0)-\boldsymbol{z}_j)}{p(\boldsymbol{x}(0))}\right]^{.5} \exp\left[-\frac{1}{8\sigma^2}\left\{\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_j)-\boldsymbol{f}^k(\boldsymbol{x}(0))]^2\right\}\right] \\
&\times \left\{\left[\sum_{k=0}^{N}\frac{\left[\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_j)-\boldsymbol{f}^k(\boldsymbol{x}(0))\right]^T}{2\sigma^2}\boldsymbol{D}\{\boldsymbol{f}^k(\boldsymbol{x}(0))\}\right] + \frac{[\boldsymbol{x}_0-\boldsymbol{x}(0)]^T}{\gamma^2}\right\}^T
\end{aligned}$$

$$\mathcal{D}_{\mathcal{A}}(j) = \left[\frac{p(\boldsymbol{x}(0)-\boldsymbol{z}_j)}{p(\boldsymbol{x}(0))}\right]^{.5} \exp\left[-\frac{1}{8\sigma^2}\left\{\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_j)-\boldsymbol{f}^k(\boldsymbol{x}(0))]^2\right\}\right] \tag{A.37}$$

$$B_{ij} = E_{\boldsymbol{x}(0)}\left\{\frac{\mathcal{N}_{\mathcal{B}}(i,j)}{\mathcal{D}_{\mathcal{B}}(i,j)}\right\} \tag{A.38}$$

$$\begin{aligned}
\mathcal{N}_{\mathcal{B}}(i,j) &= \left\{\frac{[p(\boldsymbol{x}(0)+\boldsymbol{z}_i)\,p(\boldsymbol{x}(0)+\boldsymbol{z}_j)]^{.5}}{p(\boldsymbol{x}(0))}\right. \\
&\times \left.\exp\left[-\frac{1}{8\sigma^2}\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)+\boldsymbol{z}_i)-\boldsymbol{f}^k(\boldsymbol{x}(0)+\boldsymbol{z}_j)]^2\right]\right\} \\
&+ \left\{\frac{[p(\boldsymbol{x}(0)+\boldsymbol{z}_i)\,p(\boldsymbol{x}(0)-\boldsymbol{z}_j)]^{.5}}{p(\boldsymbol{x}(0))}\right. \\
&\times \left.\exp\left[-\frac{1}{8\sigma^2}\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)+\boldsymbol{z}_i)-\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_j)]^2\right]\right\} \\
&+ \left\{\frac{[p(\boldsymbol{x}(0)-\boldsymbol{z}_i)\,p(\boldsymbol{x}(0)+\boldsymbol{z}_j)]^{.5}}{p(\boldsymbol{x}(0))}\right. \\
&\times \left.\exp\left[-\frac{1}{8\sigma^2}\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_i)-\boldsymbol{f}^k(\boldsymbol{x}(0)+\boldsymbol{z}_j)]^2\right]\right\} \\
&+ \left\{\frac{[p(\boldsymbol{x}(0)-\boldsymbol{z}_i)\,p(\boldsymbol{x}(0)-\boldsymbol{z}_j)]^{.5}}{p(\boldsymbol{x}(0))}\right. \\
&\times \left.\exp\left[-\frac{1}{8\sigma^2}\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_i)-\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_j)]^2\right]\right\}
\end{aligned} \tag{A.39}$$

$$\begin{aligned}
\mathcal{D}_{\mathcal{B}}(i,j) &= \left[\frac{p(\boldsymbol{x}(0)-\boldsymbol{z}_i)}{p(\boldsymbol{x}(0))}\right]^{.5} \exp\left[-\frac{1}{8\sigma^2}\left\{\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_i)-\boldsymbol{f}^k(\boldsymbol{x}(0))]^2\right\}\right] \\
&\times \left[\frac{p(\boldsymbol{x}(0)-\boldsymbol{z}_j)}{p(\boldsymbol{x}(0))}\right]^{.5} \exp\left[-\frac{1}{8\sigma^2}\left\{\sum_{k=0}^{N}[\boldsymbol{f}^k(\boldsymbol{x}(0)-\boldsymbol{z}_j)-\boldsymbol{f}^k(\boldsymbol{x}(0))]^2\right\}\right]
\end{aligned} \tag{A.40}$$

where $\boldsymbol{A}_{.,j}$ denotes the j^{th} column of $\boldsymbol{A}$.

Appendix B

Proofs for Chapters 6 and 7

Proof of Proposition 1: Given a matrix of rational-valued state transition probabilities $\boldsymbol{P} = [p_{ij}]_{i,j=1}^{m}$ and a row vector $\boldsymbol{\Pi}(0) = [\pi_1(0), \cdots \pi_m(0)]$ of nonzero, rational-valued, initial state probabilities, where m is the number of states, we synthesize a piecewise linear Markov map f using the procedure outlined in Section 6.3. Because each initial state probability π_j is rational-valued and equals $\lambda(I_j)$, the length of the corresponding subinterval, the endpoints of the subintervals I_j are rational-valued. Also, because each state transition probability p_{jk} is rational-valued, the slope τ_{jk} of the affine transformation which maps I_{jk} onto I_k and which is given by $\tau_{jk} = \frac{\lambda(I_k)}{p_{jk}\lambda(I_j)}$ is rational-valued as well. Because it is rational-valued, $\tau_{jk} = \frac{n_{jk}}{d_{jk}}$ where n_{jk} and d_{jk} are integers.

As in the outline of the synthesis procedure, let $(e_{jk,l}, e_{k,l})$ denote the left endpoint of this affine segment in (x, y)-coordinates, and let $(e_{jk,r}, e_{k,r})$ denote the right endpoint, when the affine segment is treated as a line segment in the (x, y)-plane. We now replace the affine transformation mapping I_{jk} onto I_k, by d_{jk} affine transformations, with the slope of each transformation equal to $d_{jk}\tau_{jk} = n_{jk}$, hence an integer, and the left and right endpoint pairs of the l^{th} such transformation given by $(e_{jk,l} + \frac{l}{d_{jk}}, e_{k,l})$ and $(e_{jk,xl} + \frac{l+1}{d_{jk}}, e_{k,r})$ for $l = 0, 1, \cdots, d_{jk} - 1$. What we in fact have done is replace the original affine transformation with d_{jk} affine transformations each of which maps a subinterval of length $\lambda(I_{jk})/d_{jk}$ onto I_k. Thus, the range of each transformation is identical to that of the original transformation. We do this to each of the affine transformations so that each transformation of the resulting map has integer slope. If we seek to guarantee that the map is an EMC map, we do a similar substitution to any affine segments with slopes of 1 or -1; we are free to specify the

number of affine segment replacements for these segments.

We must show that state sequences that arise with the transformed map are those of a Markov chain with the same TPM as that of the original map. We use a minor adaptation of the proof provided in [61] for the case in which there is a single affine transformation between pairs of states. As in the proof in [61], it is sufficient to show that for each positive, integer n

$$\lambda(I_{i_0 i_1 \cdots i_n}) = \frac{\lambda(I_{i_0 i_1})}{\lambda(I_{i_1})} \lambda(I_{i_1 i_2 \cdots i_n}) \tag{B.1}$$

where

$$I_{i_0 i_1 \cdots i_n} \equiv \{x : x \in I_{i_0}, f(x) \in I_{i_1}, \cdots, f^n(x) \in I_{i_n}\} \tag{B.2}$$

The sufficiency of the condition follows from two facts. First, by definition of a Markov process, state sequences defined on the partition $\{I_j\}$ are those of a Markov chain if the following is true:

$$\lambda(I_{i_0 i_1 \cdots i_n}) = \lambda(I_{i_0}) \frac{\lambda(I_{i_0 i_1})}{\lambda(I_{i_0})} \frac{\lambda(I_{i_1 i_2})}{\lambda(I_{i_1})} \cdots \frac{\lambda(I_{i_{n-1} i_n})}{\lambda(I_{i_{n-1}})}. \tag{B.3}$$

However, if (B.3) holds, (B.1) holds as well. Second, using an inductive argument, one can show that (B.1) implies (B.3) as well. Therefore, (B.1) and (B.3) are equivalent conditions, and thus establishing the validity of (B.1) is sufficient for establishing that state sequences that arise with the transformed map are those of a Markov chain.

To establish the validity of (B.1), we begin by inductively showing that

$$f(I_{i_0 i_1 \cdots i_n}) = I_{i_1 \cdots i_n}, \text{ almost everywhere} \tag{B.4}$$

whenever $I_{i_0 i_1 \cdots i_n}$ is nonempty. Since $f(I_{ijk}) \subset I_{jk}$, it follows that $\lambda(f(I_{ijk})) \leq \lambda(I_{jk})$. Also, since $I_{ij} = \bigcup_k I_{ijk}$, it follows that $f(I_{ij}) = \bigcup_k f(I_{ijk})$ and hence

$$I_j = f(I_{ij}) = \bigcup_k f(I_{ijk}). \tag{B.5}$$

Using the facts that $f(I_{ijk}) \neq f(I_{ijl})$ for $k \neq l$ and the restriction of f to I_{ij} consists of d_{ij} affine transformations each with slope n_{ij}, results in the following chain of equalities:

$$\lambda(I_j) = \lambda\left(\bigcup_k f(I_{ijk})\right) \tag{B.6}$$

$$= \sum_k \lambda(f(I_{ijk})) \tag{B.7}$$

$$= \sum_k n_{ij} \frac{\lambda(I_{ijk})}{d_{ij}} \tag{B.8}$$

$$= \sum_k \tau_{ij} \lambda(I_{ijk}). \tag{B.9}$$

The third equality follows from the fact that we've replaced the original affine transformation with slope τ_{ij} which maps I_{ij} onto I_j by d_{ij} affine transformations of slope n_{ij} each mapping a subinterval of I_{ij} of length $\lambda(I_{ij})/d_{ij}$ onto I_j. Since the range of each of the transformations is identical, a subinterval of I_{ijk} of length $\lambda(I_{ijk})/d_{ij}$ is in the domain of each of the d_{ij} transformations. Equivalently, the range of $f(I_{ijk})$ is the same as the range of f restricted to each of these d_{ij} subintervals of I_{ijk}, and because f is piecewise linear, $\lambda(f(I_{ijk})) = n_{ij}\frac{\lambda(f(I_{ijk}))}{d_{ij}}$ which is the length of the image of f (with slope n_{ij}) restricted to one of the subintervals.

Now $\lambda(I_j) = \sum_k \lambda(I_{jk})$ and $\lambda(f(I_{ijk})) = \tau_{ij}\lambda(I_{ijk}) \leq \lambda(I_{jk})$. Therefore,

$$\lambda(I_j) = \sum_k \lambda(I_{jk}) = \sum_k \tau_{ij}\lambda(I_{ijk}) \tag{B.10}$$

can be true only if $\lambda(I_{jk}) = \tau_{ij}\lambda(I_{ijk})$. Since f is piecewise linear on I_{ij} with each of the d_{ij} affine pieces having slope n_{ij}, a domain of length $\lambda(I_{ij})/d_{ij}$, and range I_j, this is equivalent to

$$f(I_{ijk}) = I_{jk} \text{ almost everywhere} \tag{B.11}$$

To complete the induction, we assume that for all nonempty $I_{i_0 \cdots i_l}$ the following is true

$$f(I_{i_0 \cdots i_l}) = I_{i_1 \cdots i_l} \tag{B.12}$$

for all $l \leq n-1$. Consider any nonempty set $I_{i_0 \cdots i_{n-1}}$. Then proceeding as above, we have $f(I_{i_0 \cdots i_{n-1}}) \subset I_{i_1 \cdots i_{n-1}}$, and therefore $\lambda(f(I_{i_0 \cdots i_{n-1}})) = \tau_{i_0 i_1}\lambda(I_{i_0 \cdots i_{n-1}}) \leq \lambda(I_{i_1 \cdots i_{n-1}})$. However, $I_{i_0 \cdots i_{n-1}} = \bigcup_{i_n} I_{i_0 \cdots i_{n-1} i_n}$ and thus $f(I_{i_0 \cdots i_{n-1}}) = \bigcup_{i_n} f(I_{i_0 \cdots i_{n-1} i_n})$. According to the induction hypothesis, this implies

$$I_{i_1 \cdots i_{n-1}} = \bigcup_{i_n} f(I_{i_0 \cdots i_{n-1} i_n}) \tag{B.13}$$

and thus the following is true:

$$\lambda(I_{i_1\cdots i_{n-1}}) = \sum_{i_n} n_{i_0 i_1} \frac{\lambda(I_{i_0\cdots i_{n-1} i_n})}{d_{i_0 i_1}} \tag{B.14}$$

$$= \sum_{i_n} \tau_{i_0 i_1} \lambda(I_{i_0\cdots i_{n-1} i_n}) \tag{B.15}$$

Using the fact that $I_{i_1\cdots i_{n-1}} = \bigcup_{i_n} I_{i_1\cdots i_{n-1} i_n}$ leads to the following equality

$$\sum_{i_n} \lambda(I_{i_1\cdots i_n}) = \sum_{i_n} \tau_{i_0 i_1} \lambda(I_{i_0\cdots i_n}) \tag{B.16}$$

Since no individual term in the right member of the above equation can exceed the corresponding term in the left member, the equation is valid only if

$$\lambda(I_{i_1\cdots i_n}) = \tau_{i_0 i_1} \lambda(I_{i_0\cdots i_n}) \tag{B.17}$$

Because f is piecewise linear on $I_{i_0 i_1}$ and using a similar argument as above, we must have

$$f(I_{i_0 i_1\cdots i_n}) = I_{i_1 i_2\cdots i_n} \text{ almost everywhere} \tag{B.18}$$

This completes the induction. Now observe that

$$\lambda(I_{i_0 i_1\cdots i_n}) = \frac{\lambda(I_{i_1\cdots i_n})}{\tau_{i_0 i_1}} \tag{B.19}$$

But, $\tau_{ij} = \lambda(I_j)/\lambda(I_{ij})$ whenever I_{ij} is nonempty. Therefore, B.19 may be expressed

$$\lambda(I_{i_0 i_1\cdots i_n}) = \frac{\lambda(I_{i_0 i_1})}{\lambda(I_{i_1})} \lambda(I_{i_1\cdots i_n}) \tag{B.20}$$

which is sufficient to guarantee that the state sequence is that of a Markov chain. To establish the fact that the transition probabilities are the same as those of the original map, we note that the portion of partition element I_j mapped onto partition element element I_k is given by I_{jk} for both maps. Finally, we note that the proof still holds if λ is replaced by any probability measure on the unit interval for which there is a corresponding PDF that is constant on each partition element. □

Proof of Proposition 2: Let β be the integer-valued, least common denominator of the endpoints of the affine segments and the images of these endpoints. Such a β exists since these points are all rational-valued by assumption. Consider the β element, uniform partition with the set of partition points $P = \{\frac{i}{\beta}\}_{i=0}^{\beta}$, with each partition element having length $1/\beta$. The endpoint of the affine segments and their images are all partition points since β is the least common denominator of these points. Thus, both the domain and range of each affine segment consist of connected unions of partition elements, i.e., the domain and range are subintervals.

The restriction of the map f to the domain of each affine segment is an affine transformation and thus expressible as $x(n+1) = f(x(n)) = ax(n) + b$ where a is integer-valued by assumption. The left endpoint of the affine segment and its image are given by $\frac{i}{\beta}$ and $\frac{j}{\beta}$, respectively, for some integers i and j, since they are partition points. It follows that the endpoints of the other partition elements in the domain of the affine segment are given by $\{\frac{i+k}{\beta}\}_{k=1}^{n}$ for some n. For each such endpoint $\frac{i+k}{\beta}$, we have the following:

$$f(\frac{i+k}{\beta}) = a\frac{i+k}{\beta} + b \tag{B.21}$$

$$= (a\frac{i}{\beta} + b) + \frac{ak}{\beta} \tag{B.22}$$

$$= \frac{j}{\beta} + \frac{ak}{\beta} \tag{B.23}$$

$$= \frac{j+ak}{\beta}. \tag{B.24}$$

Since j, a, k are all integer-valued, the image of $\frac{i+k}{\beta}$ is also a partition point. This condition holds for the partition points in the domain of each affine segment and thus for all partition points. Therefore, each partition point is mapped to a partition point. Also, whereas the restriction of f to the domain of each affine segment is affine and thus continuous and the image of an interval under a continuous one-dimensional mapping is also an interval, each partition element is mapped to a union of partition elements.

For the second part of the proof, let P' be the partition points for a uniform refinement of the original uniform partition. Whereas the refinement is also a uniform partition, it follows that $P' = \{\frac{i}{\gamma}\}_{i=0}^{\gamma}$ for some integer γ. Therefore, $\gamma = C\beta$ for some integer C and the partition points are expressible as $P' = \{\frac{i}{C\beta}\}_{i=0}^{C\beta}$. We can now apply the first part of the proof to this partition, since the endpoints of the affine segments are also partition points,

with each given by $\frac{C_i}{C\beta}$ for some i, and the domain of each affine segment is a union of partition elements. □

Proof of Proposition 3: a. We first show that if an EMC map f gives rise to a Markov chain with irreducible TPM, the EMC map is ergodic and each subinterval of the unit interval has nonzero (invariant) measure. If f gives rise to a Markov chain with irreducible TPM, then f is a class C function as defined in [12] and by Theorem 1 of the same reference has a unique, invariant measure that is absolutely continuous with respect to Lebesgue measure, and thus has a unique, stationary PDF p_F. Furthermore, this PDF is nonzero on each subinterval as is easily verified as follows. As shown in [12], given any Markov partition $\{I_j\}_{j=1}^N$ for a a class C function, the Frobenius-Perron operator restricted to PDFs that are constant over each partition element is a linear operator and thus can be represented by a matrix M. That is, if $p(x)$ denotes a piecewise constant PDF satisfying $p(x) = p_j$ for all $x \in I_j$ and $\boldsymbol{p} = [p_1, \cdots, p_N]$ is a row vector, $P_f(p(x))$, the Frobenius-Perron operator applied to $p(x)$, is given by $\boldsymbol{p}\,\boldsymbol{M}$. It is straightforward to show that this matrix is related to the TPM $\boldsymbol{P}$ of the Markov chain corresponding to the Markov partition by the following similarity transformation:

$$\boldsymbol{M} = \boldsymbol{D}\,\boldsymbol{P}\,\boldsymbol{D}^{-1} \tag{B.25}$$

where $\boldsymbol{D}$ is a diagonal matrix with i^{th} diagonal element given by $\lambda(I_j)$. If $\boldsymbol{P}$ is irreducible, it has a unique, invariant probability vector $\boldsymbol{\Pi}$ with no zero-valued elements. Therefore, $\boldsymbol{M}$ has a unique invariant row vector with no zero-valued elements given by $\boldsymbol{\Pi}\,\boldsymbol{D}^{-1}$. Because it is invariant, this vector corresponds to a fixed point of the Frobenius-Perron operator , and thus its elements are the constant PDF values of the partition elements for the unique, stationary PDF of f. Since f has a unique stationary PDF that is nonzero over the unit interval (except possibly at the endpoints of partition elements), f is ergodic by Theorem [50, p.55:Theorem 4.2.2].

We now show that if an EMC map f is ergodic with respect to an invariant measure μ_F with corresponding PDF p_F that is nonzero almost everywhere, each Markov chain it gives rise to has an irreducible TPM. Let $\{I_j\}_{j=1}^N$ denote the elements of any Markov partition, $\{S_j\}_{j=1}^N$ denote the states of the associated Markov chain, and $\boldsymbol{P} = [p_{ij}]$ denote the corresponding TPM. If $\boldsymbol{P} = [p_{ij}]$ is not irreducible, there exists two states S_i and S_j

such that it is impossible to ever get to S_j from S_i, or equivalently $p_{ij}^n = 0$ for all positive, integer-valued n where p_{ij}^n is the ij^{th} element of $\boldsymbol{P}^n$, the n-step TPM. However, because of the relation between states and partition elements this means that $f^n(I_i) \cap I_j = \emptyset$ for all positive, integer-valued n. Therefore, for all $x \in I_i$ and with 1_{I_j} denoting the indicator function over I_j, the following holds:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{I_j}(f^i(x)) = 0 \tag{B.26}$$

$$< \int 1_{I_j}(x)\, p_F(x)\, dx = \mu_F(1_{I_j}) \tag{B.27}$$

where the inequality holds because $\mu_F(1_{I_j}) > 0$ since $p_F(x)$ is nonzero almost everywhere. However, the inequality is a contradiction of the Birkhoff ergodic theorem. Therefore, $\boldsymbol{P}$ must be irreducible.

Note that as a consequence of the proof, it follows that for an EMC map, and more generally for an MC map, to be ergodic with respect to an invariant measure having a corresponding PDF, the PDF can only be zero over subintervals which are partition elements in Markov partitions.

b. We first show that if an EMC map f gives rise to a Markov chain with a primitive TPM, the map is exact. Let p_F denote the stationary PDF of f and μ_F denote the measure this PDF gives rise to. From [50, p.66,Theorem 4.4.1.c], it suffices to show that the following holds for each density function of μ_F (where $g \in L^1(\mu_F)$ is a density function if it is nonnegative and satisfies $\int g(x)\, d\mu_F(x) = 1$):

$$\lim_{n\to\infty} \|P_f^n(g) - 1\| \equiv \lim_{n\to\infty} \int |P_f^n(g(x)) - 1|\, d\mu_F(x) = 0 \tag{B.28}$$

where P_f^n is the Frobenius-Perron operator of f^n. The above condition means that the Frobenius-Perron operator applied to any density function (with respect to μ_F) converges strongly to the constant 1. Intuitively, this means that any initial PDF for the map f converges under the dynamics of f to the stationary PDF. As noted in [50, p. 69], it suffices to prove the above condition for all g in a linearly dense subset of the set of density functions. Because μ_F is absolutely continuous with respect to Lebesgue measure on the unit interval, a linearly dense subset of density functions consists of all normalized characteristic functions $k_\beta(x)$, where $k_\beta(x) = \frac{1}{\mu_F(\beta)}$ if $x \in \beta$ and 0 otherwise and β is a subinterval of the unit

interval.

Two facts which simplify the proof follow directly from more general results in [24]. First, any MC map which gives rise to a Markov chain with primitive TPM has a dense set of eventually periodic points. Second, if an MC map gives rise to a Markov chain with primitive TPM, then all Markov chains which the map gives rise to have primitive TPMs as well. In light of the first fact, f gives rise to arbitrarily fine Markov partitions since one can find a Markov partition which includes any given eventually periodic point x. In particular, one starts with any Markov partition and uses the finite set of distinct points $\{f^i(x)\}$ as additional partition points. In addition, the set of subintervals with eventually periodic points of f as endpoints are dense in the set of all subintervals of the unit interval. Therefore, to verify (B.28) for all density functions, it suffices to verify it for normalized characteristic functions $k_\alpha(x)$, where α a subinterval with eventually periodic points of f as endpoints.

The following lemma follows as a consequence of [12, Theorem 3] which establishes the relation used earlier between the TPM of the Markov chain which arises from a Markov partition and the Frobenius-Perron operator restricted to PDFs that are constant over each partition element.

Lemma: The time evolution under the dynamics of f of any piecewise constant PDF, where the endpoints of the subintervals with constant PDF values are eventually periodic points of f, is uniquely determined by the time evolution of the TPM of a Markov chain.

Sketch of Proof: Because the endpoints are eventually periodic points, one can find a Markov partition with these endpoints among the partition points; and the PDF is constant over partition elements. Therefore, from [12, Theorem 3] and as discussed earlier the Frobenius-Perron operator restricted to PDFs that are piecewise constant over partition elements has a matrix representation which is related to the TPM of the corresponding Markov chain by a similarity transformation involving a diagonal matrix (with the length of the partition elements as the diagonal terms). Therefore, the time evolution of a piecewise constant PDF under the dynamics of f is uniquely determined by the time evolution of the TPM of the Markov chain. □

Since f gives rise to a Markov chain with primitive TPM, it has a unique, stationary density that is nonzero almost everywhere (from part a.). Now given any $k_\alpha(x)$, one can

find a Markov partition with the endpoints of α as partition points. Because the TPM $\boldsymbol{P}$ of the corresponding Markov chain is primitive, any vector of initial state probabilities $\boldsymbol{\Pi}(0)$ converges to the unique, invariant probability vector, which we denote $\boldsymbol{\Pi}$, of the TPM. Let $\boldsymbol{\Pi}(0)$ be chosen such that $\pi_j(0)$, the initial probability for the state corresponding to partition element I_j, is given by

$$\pi_j(0) = \int_{I_j} k_\alpha(x)dx = \frac{\lambda(I_j \cap \alpha)}{\lambda(\alpha)}. \tag{B.29}$$

As defined, $\boldsymbol{\Pi}(0)$ is the probability vector corresponding to the density function $k_\alpha(x)$ for the chosen Markov partition. In addition, we know from the discussion in part (a.) that the Frobenius-Perron operator restricted to PDFs that are piecewise constant over this Markov partition has the matrix representation $\boldsymbol{M}$ given by (B.25). Since $\boldsymbol{\Pi}(0)$ converges to $\boldsymbol{\Pi}$, then $k_\alpha(x)$ converges to $\boldsymbol{\Pi}\,\boldsymbol{D}^{-1}$, where $\boldsymbol{D}$ is a diagonal matrix with j^{th} diagonal element given by $\lambda(I_j)$. However, by the uniqueness of the stationary density of f, $\boldsymbol{\Pi}\,\boldsymbol{D}^{-1}$ must equal this density. Therefore, the initial PDF $k_\alpha(x)$ converges pointwise to the unique stationary PDF of f. The convergence is also convergence in $L^1(\mu_F)$, i.e., (B.28) holds with $g(x) = k_\alpha(x)$, because $P_f^n(k_\alpha)$ is piecewise constant for each n, with a finite number of pieces no greater than the number of elements in the chosen Markov partition. Since α is arbitrary, the result holds for each subinterval with endpoints given by eventually periodic points of f. Therefore, the result holds for all density functions.

We now show that if f is exact with respect to an invariant measure μ that is nonzero over every subinterval of the unit interval, then the TPM of each Markov chain it gives rise to is primitive. Consider any Markov partition $\{I_j\}$ for f and let S_j denote the state associated with I_j of the corresponding Markov chain. Since f is exact

$$\lim_{n\to\infty} \mu(f^n(I_j)) = 1. \tag{B.30}$$

However, since I_j is a partition element, $f(I_j)$ is a union of partition elements. Furthermore, each partition element has finite measure and the number of partition elements is finite. Therefore, the exactness property must be satisfied for some finite integer $N(j)$, i.e., $\mu(f^{N(j)}(I_j)) = 1$, and it must also be true that $f^{N(j)}(I_j) = I$ (except possibly on a set of measure zero), where I is the unit interval. In light of the correspondence between partition elements and states of the Markov chain, it follows that each state of the Markov chain is

taken to every state of the chain after N time steps where N is the largest of the $N(j)$. Therefore, the TPM of the Markov chain is primitive. □

Proof of Proposition 4: As noted in the discussion before the statement of the proposition, if the sequence of affine parameter pairs $\{(\tau(i,\hat{x}_{ML}),\beta(i,\hat{x}_{ML}))\}_{i=0}^{N}$ is known, then one can determine the ML orbit segment $\{f^i(\hat{x}_{ML})\}_{i=0}^{N}$. As a consequence of the following lemma, one can determine this sequence by exploiting the relation between noise-corrupted orbit segments of MC maps and hidden Markov models (HMMs).

Lemma: For any MC map which gives rise to arbitrarily fine Markov partitions and any finite set of observations $Y = \{y(i)\}_{i=0}^{N}$, one can find a Markov partition for which the corresponding Markov chain has the property that for an appropriately defined HMM (with the definition not dependent on the observation sequence), the sequence of affine parameter pairs associated with the most likely state sequence S_{ML} is identical to the sequence of affine parameter pairs associated with $\hat{x}_{ML}$. (The definition of S_{ML} and the sequences of affine parameter pairs associated with S_{ML} and the $\hat{x}_{ML}$ are provided in the discussion before the statement of the proposition).

Proof of Lemma: As in Section 7.6, let L denote the minimum number of affine segments of f, (τ_i,β_i) denote the pair of affine parameters associated with the i^{th} affine segment, and A_i denote the domain of this segment, so that $f(x) = \tau_i x + \beta_i$ if $x \in A_i$. As noted in the section, one can associate a sequence of $N+1$ affine segment domains, $\{A(i,x)\}_{i=0}^{N}$ with each initial condition x, and this sequence is associated with a subinterval of initial conditions $A(x)$ which is given by

$$A(x) = \bigcap_{i=0}^{N} f^{-i}(A(i,x)) \tag{B.31}$$

where f^{-i} denotes the inverse image of the composed map f^i. One can show that two such subintervals $A(x)$ and $A(y)$ are either identical or disjoint and these subintervals form a partition of the unit interval, which is in fact a Markov partition. Thus the subintervals form a set of equivalence classes of points on the unit interval. Let $A_{EQ} = \{A(x_i)\}$ denote a complete set of these equivalence classes, i.e., the $\{A(x_i)\}$ are disjoint subintervals whose union equals the unit interval. There are L^{N+1} sequences of affine domain segments not all of which are associated with initial conditions. As such there are at most L^{N+1} equivalence

classes in A_{EQ}. It follows from the discussion in the section that the restriction of the log-likelihood function (7.50) to each equivalence class in A_{EQ} (when treated as a function of the initial condition x) is a quadratic function in x since the same sequence of $N+1$ affine parameter pairs is associated with each point in an equivalence class.

Let $LF(A(x_i))$ denote the supremum (generally maximum) value of the log-likelihood function $\log p(Y;x)$ given by (7.50) restricted to $x \in A(x_i)$. By definition, the log-likelihood function attains its maximum value on the unit interval at $\hat{x}_{ML}$. It follows that this value is also the largest of the $\{LF(A(x_i))\}$ and is associated with that subinterval $A(x_j)$ which contains $\hat{x}_{ML}$, so that $A(\hat{x}_{ML}) = A(x_j)$. Let δ_{MIN} denote the difference between $LF(A(\hat{x}_{ML}))$ and the next largest of the $\{LF(A(x_i))\}$. In other words, δ_{MIN} represents the smallest difference between the value of the likelihood function for $\hat{x}_{ML}$ and the value of the likelihood function for all other initial conditions on the unit interval with associated sequences of affine segment domains which differ from the sequence associated with $\hat{x}_{ML}$.

Now consider any Markov partition for f and its associated Markov chain. Also, consider the HMM model introduced in Section 7.6 which uses the state transition pseudo-probabilities $\{q_{ij}\}$, the initial state pseudo-probabilities $\eta(S_j) = 1$, and for which the output PDF $p(y|S_j)$ associated with each state S_j is given by

$$p(y|S_j) = \frac{\exp\left[-\frac{(y-H_j)^2}{2\sigma^2}\right]}{(2\pi\sigma^2)^{1/2}}. \tag{B.32}$$

where $H_j \in I_j$ and I_j is the partition element associated with S_j.

A fact which follows from the discussion in the chapter is that there is a state sequence $S^x = \{S^x(i)\}_{i=0}^N$ associated with each initial condition x, where $S^x(i) \equiv S_j$ if $f^i(x) \in I_j$. In addition, given any state sequence $S = \{S(i)\}_{i=0}^N$ with nonzero probability, i.e., $P(S) > 0$, there is an initial condition (actually, a subinterval of initial conditions) x associated with S, in the sense that $S^x = S$ which means that $S^x(i) = S(i)$ for $0 \le i \le N$. It follows from this that for any state sequence S for which $q(S,Y) > 0$, where $q(S,Y)$ is the joint pseudo-likelihood of the state sequence and observation set, there exists some initial condition x such that $S^x = S$.

For arbitrary initial condition x, we now derive an upper bound on the absolute difference $|q(S^x,Y) - p(Y;x)|$, which can be thought of not only as the absolute difference of the joint pseudo-likelihood of the state sequence associated with x and the likelihood of x, but also as

the absolute difference of the joint pseudo-likelihood of a state sequence and the likelihood of any initial condition associated with that state sequence. In light of the close relation between $q(S^x, Y)$ and the log-likelihood function (7.34), what we in fact are upper bounding is the difference between the value of the likelihood function for the pseudo-orbit segment $\{H(S^x(i))\}_{i=0}^N$ and any actual orbit segment with associated state sequence S^x (where $H(S^x(i)) = H_j$ if $S^x(i) = S_j$). We now show that such an upper bound is provided by γ_{MAX} which is given by

$$\gamma_{MAX} \equiv \frac{1 - \left(1 - \frac{\epsilon \exp[-\frac{1}{2}]}{\sigma}\right)^{N+1}}{(2\pi\sigma^2)^{(N+1)/2}}. \tag{B.33}$$

where ϵ is the length of the longest partition element.

To establish the validity of (B.33), we first note that for a fixed y, the Gaussian PDF $p(y;x)$ given by

$$p(y;x) = \frac{\exp\left[-\frac{(y-x)^2}{2\sigma^2}\right]}{(2\pi\sigma^2)^{1/2}} \tag{B.34}$$

has a slope attaining its maximum absolute value at $x = y \pm \sigma$, with this maximum absolute value p'_{MAX} given by

$$p'_{MAX} = \frac{\exp[-\frac{1}{2}]}{(2\pi\sigma^4)^{1/2}}. \tag{B.35}$$

Therefore, for any two points x_1 and x_2 with $|x_1 - x_2| < \epsilon$, the absolute difference $|p(y;x_1) - p(y;x_2)|$ is upper bounded by $p'_{MAX}\,\epsilon$. In addition, one can show that for positive real numbers a, b, and c with $a > b$, and positive integer n, $[(a+c)^n - (b+c)^n] > [a^n - b^n]$. Since the maximum value of $p(y;x)$ is $(2\pi\sigma^2)^{-1/2}$ (attained at $x = y$), it follows that for a fixed sequence $\{y(i)\}_{i=0}^N$ and variable sequences $\{x_1(i)\}_{i=0}^N$ and $\{x_2(i)\}_{i=0}^N$ satisfying $|x_1(i) - x_2(i)| < \epsilon$ for each i, an upper bound on the absolute difference $|\prod_{i=0}^N p(y(i)|x_1(i)) - \prod_{j=0}^N p(y(j)|x_2(j))|$ is given by

$$|\prod_{i=0}^{N} p(y(i)|x_1(i)) - \prod_{j=0}^{N} p(y(j)|x_2(j))| \;<\; \frac{1 - (1 - p'_{MAX}\,\epsilon)^{N+1}}{(2\pi\sigma^2)^{(N+1)/2}} \tag{B.36}$$

$$= \frac{1 - \left(1 - \frac{\epsilon \exp[-\frac{1}{2}]}{\sigma}\right)^{N+1}}{(2\pi\sigma^2)^{(N+1)/2}}. \tag{B.37}$$

However, the product $\prod_{i=0}^N p(y(i);x_1(i))$ is the likelihood of a state sequence if for each

i, $x_1(i) = H_j$ for some j. Similarly, the product $\prod_{j=0}^{N} p(y(j); x_2(j))$ is the value of the likelihood function when $\{x_2(i)\}$ is an orbit segment. Therefore, (B.37) provides an upper bound on $|q(S^x, Y) - p(Y; x)|$ as well.

Because γ_{MAX} is an increasing function of ϵ and f gives rise to arbitrarily fine Markov partitions (by assumption), we can find a Markov partition with small enough ϵ such that

$$2\gamma_{MAX} < \delta_{MIN}. \tag{B.38}$$

where δ_{MIN} was defined earlier as the smallest difference between the value of the likelihood function for $\hat{x}_{ML}$ and the value of the likelihood function for all other initial conditions on the unit interval with associated sequences of affine segment domains which differ from the sequence associated with $\hat{x}_{ML}$.

Given such a Markov partition and the Markov chain corresponding to the partition, consider $q(S^{\hat{x}_{ML}}, Y)$, the joint likelihood of the observation set and state sequence associated with $\hat{x}_{ML}$. Because $S^{\hat{x}_{ML}}$ is the state sequence associated with $\hat{x}_{ML}$ and (B.38) holds, it follows that

$$\begin{aligned} q(S^{\hat{x}_{ML}}, Y) &> p(Y; \hat{x}_{ML}) - \gamma_{MAX} && \text{(B.39)} \\ &> p(Y; \hat{x}_{ML}) - \frac{\delta_{MIN}}{2}. && \text{(B.40)} \end{aligned}$$

Now let S_{BIG} denote the state sequence which maximizes the joint likelihood $q(S, Y)$ among those state sequences for which the associated sequence of affine parameter pairs is different from the sequence associated with $S^{\hat{x}_{ML}}$. As noted earlier, if $q(S_{BIG}, Y) > 0$ then S_{BIG} is associated with at least one initial condition z of f. However, because the same sequence of affine parameter pairs is associated with both S_{BIG} and z and this sequence is different from the sequence associated with $\hat{x}_{ML}$, the following must be true

$$p(Y; z) \leq p(Y; \hat{x}_{ML}) - \delta_{MIN} \tag{B.41}$$

by definition of δ_{MIN}. Therefore, we have the following chain of inequalities:

$$\begin{aligned} q(S_{BIG}, Y) &< p(Y; z) + \gamma_{MAX} && \text{(B.42)} \\ &\leq p(Y; \hat{x}_{ML}) - \delta_{MIN} + \gamma_{MAX} && \text{(B.43)} \end{aligned}$$

$$< \quad p(Y;\hat{x}_{ML}) - \delta_{MIN} + \frac{\delta_{MIN}}{2} \tag{B.44}$$

$$= \quad p(Y;\hat{x}_{ML}) - \frac{\delta_{MIN}}{2} \tag{B.45}$$

Combining (B.40) and (B.45) yields the following:

$$q(S^{\hat{x}_{ML}},Y) \quad > \quad p(Y|\hat{x}_{ML}) - \frac{\delta_{MIN}}{2} \tag{B.46}$$

$$> \quad q(S_{BIG},Y) \tag{B.47}$$

The inequality means that with the chosen Markov partition and with any refinement of the partition, the state sequence associated with $\hat{x}_{ML}$ has a greater joint likelihood, or equivalently a larger likelihood, than the likelihood of any other state sequence for which the sequence of associated affine parameter pairs differs from that associated with $\hat{x}_{ML}$. Therefore, the sequence of affine parameter pairs associated with the state sequence with largest likelihood must be identical to that associated with $\hat{x}_{ML}$.

An important observation is that this result does not imply that $S^{\hat{x}_{ML}}$ is the most likely state sequence. The sequence of affine parameter pairs associated with $S^{\hat{x}_{ML}}$ is associated with other state sequences as well, and one of these state sequences may be the most likely state sequence. □

In light of the lemma, one can find the sequence of affine parameter pairs associated with $\hat{x}_{ML}$ by first finding a sufficiently fine Markov partition and its corresponding Markov chain and then determining the most likely state sequence for an appropriately defined HMM and the sequence of affine parameter pairs associated with this state sequence. If the partition is fine enough, this sequence of affine parameter pairs will be the same as the sequence associated with $\hat{x}_{ML}$. □

Bibliography

[1] H. Abarbanel, R. Brown, and M. Kennel, "Lyapunov Exponents in Chaotic Systems: Their Importance and Their Evaluation Using Observed Data," *International Journal of Modern Physics B*, Vol. 5, No. 9, 1991, pp. 1347–1375.

[2] H. Abarbanel, R. Brown, and M. Kennel, "Variation of Local Lyapunov Exponents on a Strange Attractor," *Journal of Nonlinear Science*, Vol. 1, 1991, pp. 175–199.

[3] R. Adler and B. Weiss, "Similarity of Automorphisms of the Torus," *Memoirs of the American Mathematical Society*, Vol. 98, 1970.

[4] D. Allspach and H. Sorenson, "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," *IEEE Transactions on Automatic Control*, Vol. AC-17, No. 4, 1972, pp. 439–448.

[5] B. Anderson and J. Moore, *Optimal Filtering*, Englewood Cliffs, NJ: Prentice Hall, 1979.

[6] E. Barankin, "Locally Best Unbiased Estimates," *Annals of Mathematical Statistics*, Vol. 20, 1949, pp. 477–501.

[7] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice Hall, 1979.

[8] M. Bertero, T. Poggio, and V. Torre, "Ill-Posed Problems in Early Vision," *Proceedings of the IEEE*, Vol. 76, No. 8, August 1988, pp. 869–889.

[9] B. Bobrovsky and M. Zakai, "A Lower Bound on the Estimation Error for Certain Diffusion Processes," *IEEE Transactions on Information Theory*, Vol. IT-22, No. 1, 1976, pp. 45–52.

[10] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, New York: Springer-Verlag, 1975.

[11] R. Bowen, *On Axiom-A Diffeomorphisms*, Regional Conference Series in Mathematics, No. 35, Providence: American Mathematical Society, 1978.

[12] A. Boyarsky and M. Scarowsky, "On a Class of Transformations which have Unique Absolutely Continuous Invariant Measures," *Transactions of the American Mathematical Society*, Volume 255, November 1979, pp. 243–262.

[13] A. Boyarsky, "Randomness Implies Order," *Journal of Mathematical Analysis and Applications*, Vol. 76, 1980, pp. 483–496.

[14] A. Boyarsky, "A Matrix Method for Estimating the Liapunov Exponent of One-Dimensional Systems," *Journal of Statistical Physics*, Vol. 50, Nos. 1–2, 1988, pp. 213-229.

[15] R. Bucy and K. Senne, "Digital Synthesis of Nonlinear Filters," *Automatica*, Volume 7, 1971, pp. 287–298.

[16] J. Center, Jr., "Practical Nonlinear Filtering of Discrete Observations by Generalized Least Squares Approximation of the Conditional Probability Distribution," *Proceedings of the Second Symposium on Nonlinear Estimation Theory and its Application*, North Hollywood: Western Periodicals, 1972, pp. 88–99.

[17] D. Chapman and H. Robbins, "Minimum Variance Estimation Without Regularity Assumptions," *Annals of Mathematical Statistics*, Vol. 22, 1951, pp. 581–586.

[18] K. Cheok, H. Hu, and N. Loh, "Modeling and Identification of a Class of Servomechanism Systems with Stick-Slip Friction," *Transactions of the American Society of Mechanical Engineers*, Vol. 110, September 1988, pp. 324–328.

[19] T. Cover and J. Thomas, *Elements of Information Theory*, New York: John Wiley and Sons, 1991.

[20] R. Devaney, *An Introduction to Chaotic Dynamical Systems*, Redwood City: Addison-Wesley, 1989.

[21] J. Eckmann and D. Ruelle, "Ergodic Theory of Chaos and Strange Attractors," *Reviews of Modern Physics*, Volume 57, Number 3, Part 1, July 1985, pp. 617–656.

[22] J. Farmer and J. Sidorowich, "Exploiting Chaos to Predict the Future and Reduce Noise," in *Evolution, Learning, and Cognition*, Y.C. Lee, Ed., World Scientific, 1988, pp. 277–330.

[23] J. Farmer and J. Sidorowich, "Optimal Shadowing and Noise Reduction," *Physica D*, Vol. 47, No. 3, January 1991, pp. 373–392.

[24] N. Friedman and A. Boyarsky, "Irreducibility and Primitivity Using Markov Maps," *Linear Algebra and its Applications*, Vol. 37, 1981, pp. 103–117.

[25] N. Friedman and A. Boyarsky, "Matrices and Eigenfunctions Induced by Markov Maps," *Linear Algebra and its Applications*, Vol. 38, 1981, pp. 141–147.

[26] N. Friedman and A. Boyarsky, "Irreducibility and Primitivity Using Markov Maps," *Linear Algebra and its Applications*, Vol. 37, 1981, pp. 103–117.

[27] W. Fuller, *Measurement Error Models*, New York: Wiley, 1987.

[28] J. Galdos, "A Rate Distortion Theory Lower Bound on Desired Filtering Error," *IEEE Transactions on Information Theory*, Vol. IT-27, No. 3, May 1981, pp. 366–368.

[29] F. Gantmacher, *The Theory of Matrices*, New York: Chelsea Publishing Company, 1960.

[30] A. Gelb, Ed., *Applied Optimal Estimation*, Cambridge, MA: MIT Press, 1974.

[31] L. Gleser, "Estimation in a Multivariate 'Errors in Variables' Regression Model: Large Sample Results," *The Annals of Science*, Vol. 9, No. 1, pp. 24–44.

[32] G. Golub and C. Van Loan, *Matrix Computations*, Baltimore: John Hopkins University Press, 1984.

[33] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, New York: Springer-Verlag, 1986.

[34] P. Halmos, *Lectures on Ergodic Theory*, Tokyo: The Mathematical Society of Japan, 1956.

[35] S. Hammel, "A Noise Reduction Method for Chaotic Systems," *Physics Letters A*, Vol. 148, Nos. 8–9, 3 September 1990, pp. 421–428.

[36] C. Hecht, "Digital Realization of Nonlinear Filters, *Proceedings of the Second Symposium on Nonlinear Estimation Theory and its Application*, North Hollywood: Western Periodicals, 1972, pp. 152–158.

[37] S. Isabelle, Ph. D. Thesis, Massachusetts Institute of Technology, 1993.

[38] A. Jazwinski, *Stochastic Processes and Filtering Theory*, New York: Academic Press, 1970.

[39] B. Juang and L. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models," *AT&T Technical Journal*, Vol. 64, No. 2, February 1985, pp. 391–408.

[40] J. Kadtke and J. Brush, "Noise Reduction Methods for Chaotic Signals Using Empirical Equations of Motion," *Proceedings of the SPIE Conference on Signal Processing, Sensor Fusion, and Target Recognition*, Vol. 1699, 1992, pp. 338–349.

[41] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communication Technology*, Vol. COM-15, No. 1, February 1967, pp. 52–60.

[42] G. Kallianpur, *Stochastic Filtering Theory*, New York: Springer-Verlag, 1980

[43] R. Kalman, "Nonlinear Aspects of Sampled-Data Control Systems," in *Proceedings of the Symposium on Nonlinear Circuit Analysis*, Polytechnic Institute of Brooklyn, 25-27 April 1956, pp. 273–313.

[44] J. Kiefer, "On Minimum Variance Estimators," *Annals of Mathematical Statistics*, Vol. 23, 1952, pp. 627–629.

[45] E. Kostelich and J. Yorke, "Noise Reduction: Finding the Simplest Dynamical System Consistent with the Data," *Physica D*, Vol. 41, 1990, pp. 183–196.

[46] A. Kosyakin and E. Sandler, *IZV Matematika*, Vol. 3, No. 118, 1972, pp. 32–40.

[47] H. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, New York: Academic Press, 1977.

[48] H. Kushner, "A Robust Discrete State Approximation to the Optimal Nonlinear Filter for a Diffusion," *Stochastics*, vol. 3, 1979, pp. 75–83.

[49] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Orlando: Academic Press, 1985.

[50] A. Lasota and M. Mackey, *Probabilistic Properties of Deterministic Systems*, Cambridge: Cambridge University Press, 1985.

[51] R. Larson and J. Peschon, "A Dynamic Programming Approach to Trajectory Estimation," *IEEE Transactions on Automatic Control*, Vol. AC-11, No. 3, pp. 537–540.

[52] E. Lehmann, *Theory of Point Estimation*, New York: John Wiley and Sons, 1983.

[53] L. Liporace, "Maximum Likelihood Estimation of Multivariate Observations of Markov Sources," *IEEE Transactions on Information Theory*, Vol. IT-28, No. 5, September 1982, pp. 729–734.

[54] R. Liptser and A. Shiryayev, *Statistics of Random Processes I: General Theory*, Vol. 1, New York: Springer-Verlag, 1977.

[55] M. Mackey, "The Dynamic Origin of Increasing Entropy," *Review of Modern Physics*, Vol. 61, No. 4, October 1989, pp. 981–1015.

[56] P. Marteau and H. Abarbanel, "Noise Reduction in Time Series Using Scaled Probabilistic Models," *Journal of Nonlinear Science*, Vol. 1, 1991, pp. 313–343.

[57] P. Maybeck, *Stochastic Models, Estimation, and Control, Volume 2.* New York: Academic Press, 1982.

[58] R. McAulay and L. Seidman, "A Useful Form of the Barankin Lower Bound and its Application to PPM Threshold Analysis," *IEEE Transactions on Information Theory*, Vol. IT-15, No. 2, March 1969, pp. 273–279.

[59] R. McAulay and E. Hofstetter, "Barankin Bounds on Parameter Estimation," *IEEE Transactions on Information Theory*, Vol. IT-17, No. 6, November 1971, pp. 669–676.

[60] F. Moon, *Chaotic Vibrations*, New York: John Wiley and Sons, 1987.

[61] J. Moore and S. Sengupta, "Existence and Control of Markov Chains in Systems of Deterministic Motion, *SIAM Journal of Control*, Vol. 13, No. 5, August 1975, pp. 1103–1114.

[62] J. Munkres, *Topology: A First Course*, Englewood Cliffs, NJ: Prentice-Hall, 1975.

[63] C. Myers, S. Kay, and M. Richard, "Signal Separation for Nonlinear Dynamical Systems," *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, March 1992.

[64] C. Myers, A. Singer, B. Shin, and E. Church, "Modeling Chaotic Systems with Hidden Markov Models," *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, March 1992.

[65] H. Papadopulous, *Detection and Estimation of a Class of Chaotic Signals with Application to Communications*, S.M. Thesis, Massachusetts Institute of Technology, May 1993.

[66] H. Papadopoulos and G. Wornell. "Optimal Detection of Chaotic Signals," *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, April 1993.

[67] H. Papadopoulos and G. Wornell, "Maximum Likelihood Estimation of a Class of Chaotic Signals," *IEEE Transactions on Information Theory*, 1993, To appear.

[68] A. Papoulis, *Probability and Statistics*, Englewood Cliffs, NJ: Prentice-Hall, 1990.

[69] T. Parker and L. Chua, *Practical Numerical Algorithms for Chaotic Systems*, New York: Springer-Verlag, 1989.

[70] T. Petrie, "Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, Vol. 40, No. 1, 1969, pp. 97–115.

[71] R. Pickholtz, D. Schilling, L. Milstein, "Theory of Spread-Spectrum Communications," *IEEE Transactions on Communications*, Vol. COM-30, No. 5, May 1982, pp. 855–884.

[72] V. Poor, *An Introduction to Signal Detection and Estimation*, New York: Springer-Verlag, 1988.

[73] J. Proakis, *Digital Communications*, New York: McGraw-Hill, 1989.

[74] L. Rabiner and B. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, January 1986, pp. 4–16.

[75] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications," *Proceedings of the IEEE*, Vol. 77, No. 2, February 1989, pp. 257–286.

[76] M. Richard, *Probabilistic State Estimation with Discrete-Time Chaotic Systems*, MIT Research Laboratory of Electronics Technical Report 571, March 1992.

[77] M. Richard, "State Estimation with Chaotic Maps," presented at the 1992 IEEE Digital Signal Processing Workshop, September 1992, Utica, Illinois.

[78] M. Scarowsky, A. Boyarsky, and H. Proppe, "Some Properties of Piecewise Linear Expanding Maps", *Nonlinear Analysis, Theory, Methods, and Applications*, Vol. 4, No. 1, 1980, pp. 109–121.

[79] L. Seidman, "Performance Limitations and Error Calculations for Parameter Estimation," *Proceedings of the IEEE*, Vol. 58, No. 5, May 1970, pp. 644–652.

[80] S. Sengupta, P. Czarny, and R. Chow, "A Representation Theorem for Finite Markov Chains Whose States are Subintervals of [0,1]", *Information Sciences*, Volume 3, 1971, pp. 51–58.

[81] A. Singer, "Codebook Prediction: A Nonlinear Signal Modeling Paradigm," *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, March 1992.

[82] S. Smale, "Differentiable Dynamical Systems," in *The Mathematics of Time*, New York: Springer-Verlag, 1980, pp. 1–83.

[83] H. Sorenson, "Estimation for Dynamic Systems: a Perspective," *Proceedings of the Fourth Symposium on Nonlinear Estimation Theory and its Application.* North Hollywood: Western Periodicals, 1972, pp. 291–318.

[84] H. Sorenson, "Recursive Estimation for Nonlinear Dynamic Systems," in *Bayesian Analysis of Time Series and Dynamic Models*, J. Spall, Ed., New York: Marcel Dekker, 1988, pp. 127–165.

[85] J. Taylor, "The Cramer-Rao Estimation Lower Bound Computation for Deterministic Nonlinear Systems," *IEEE Transactions on Automatic Control*, Vol. AC-24, 1979, pp. 343–344.

[86] H. Van Trees, *Detection, Estimation, and Modulation Theory, Part 1*, New York: John Wiley and Sons, 1968.

[87] P. Walters, *An Introduction to Ergodic Theory*, New York: Springer-Verlag, 1982.

[88] E. Weinstein and A. Weiss, "Lower Bounds on the Mean-Square Estimation Error," *Proceedings of the IEEE*, Vol. 73, No. 9, September 1985, pp. 1433–1434.

[89] E. Weinstein and A. Weiss, "A General Class of Lower Bounds in Parameter Estimation," *IEEE Transactions on Information Theory*, Vol. IT-34, No. 2, March 1988, pp. 338–342.

[90] A. Weiss and E. Weinstein, "A Lower Bound on the Mean-Square Error in Random Parameter Estimation," *IEEE Transactions on Information Theory*, Vol. IT-31, No. 5, September 1985, pp. 680–682.

[91] T. Yamamoto, "On the Extreme Values of the Roots of Matrices," *Journal of the Mathematical Society of Japan*, Volume 19, 1967, pp. 175–178.

[92] M. Zakai and J. Ziv, "Lower and Upper Bounds on the Optimal Filtering Error for Certain Diffusion Processes", *IEEE Transactions on Information Theory*, Vol. IT-18, No. 3, May 1972, pp. 325–331.